Real and Abstract Analysis

Kenneth Kuttler klkuttler@gmail.com

August 6, 2023

Contents

1 Review of Some Linear Algebra **15**

1.1 The Matrix of a Linear Map . . . 15
1.2 Block Multiplication of Matrices . . . 15
1.3 Schur's Theorem . . . 18
1.4 Hermitian and Symmetric Matrices . . . 20
1.5 The Right Polar Factorization . . . 21
1.6 Elementary matrices . . . 24
1.7 The Row Reduced Echelon Form Of A Matrix . . . 32
1.8 Finding the Inverse of a Matrix . . . 35
1.9 The Mathematical Theory of Determinants . . . 39
1.9.1 The Function sgn . . . 39
1.9.2 The Definition of the Determinant . . . 41
1.9.3 A Symmetric Definition . . . 43
1.9.4 Basic Properties of the Determinant . . . 44
1.9.5 Expansion Using Cofactors . . . 46
1.9.6 A Formula for the Inverse . . . 48
1.9.7 Cramer's Rule . . . 49
1.9.8 Rank of a Matrix . . . 50
1.9.9 An Identity of Cauchy . . . 52
1.10 The Cayley Hamilton Theorem . . . 53

I Topology, Continuity, Algebra, Derivatives **55**

2 Some Basic Topics **57**

2.1 Basic Definitions . . . 57
2.2 The Schroder Bernstein Theorem . . . 59
2.3 Equivalence Relations . . . 63
2.4 sup and inf . . . 63
2.5 Double Series . . . 64
2.6 lim sup and lim inf . . . 66
2.7 Nested Interval Lemma . . . 68
2.8 The Hausdorff Maximal Theorem . . . 68

3 Metric Spaces **71**

3.1 Open and Closed Sets, Sequences, Limit Points . . . 71
3.2 Cauchy Sequences, Completeness . . . 73

3.3 Closure of a Set . . . 74
3.4 Separable Metric Spaces . . . 75
3.5 Compact Sets . . . 76
3.6 Continuous Functions . . . 80
3.7 Continuity and Compactness . . . 81
3.8 Lipschitz Continuity and Contraction Maps . . . 82
3.9 Convergence of Functions . . . 84
3.10 Compactness in $C(X,Y)$ Ascoli Arzela Theorem . . . 85
3.11 Connected Sets . . . 88
3.12 Partitions of Unity in Metric Space . . . 91
3.13 Exercises . . . 93

4 Linear Spaces **99**
4.1 Algebra in $\mathbb{F}^n$, Vector Spaces . . . 99
4.2 Subspaces Spans and Bases . . . 100
4.3 Inner Product and Normed Linear Spaces . . . 103
4.3.1 The Inner Product in $\mathbb{F}^n$ . . . 103
4.3.2 General Inner Product Spaces . . . 104
4.3.3 Normed Vector Spaces . . . 106
4.3.4 The p Norms . . . 106
4.3.5 Orthonormal Bases . . . 108
4.4 Equivalence of Norms . . . 109
4.5 Covering Theorems . . . 113
4.5.1 Vitali Covering Theorem . . . 113
4.5.2 Besicovitch Covering Theorem . . . 115
4.6 Exercises . . . 120

5 Functions on Normed Linear Spaces **125**
5.1 $\mathscr{L}(V,W)$ as a Vector Space . . . 125
5.2 The Norm of a Linear Map, Operator Norm . . . 126
5.3 Continuous Functions in Normed Linear Space . . . 128
5.4 Polynomials . . . 129
5.5 Weierstrass Approximation Theorem . . . 130
5.6 Functions of Many Variables . . . 132
5.7 A Generalization . . . 134
5.8 An Approach to the Integral . . . 136
5.9 The Stone Weierstrass Approximation Theorem . . . 141
5.10 Connectedness in Normed Linear Space . . . 144
5.11 Saddle Points* . . . 145
5.12 Exercises . . . 150

6 Fixed Point Theorems **157**
6.1 Simplices and Triangulations . . . 157
6.2 Labeling Vertices . . . 161
6.3 The Brouwer Fixed Point Theorem . . . 163
6.4 The Schauder Fixed Point Theorem . . . 165
6.5 The Kakutani Fixed Point Theorem . . . 169
6.6 Ekeland's Variational Principle . . . 172

6.6.1 Cariste Fixed Point Theorem . . . 174
6.6.2 A Density Result . . . 175
6.7 Exercises . . . 177

7 The Derivative 185
7.1 Limits of a Function . . . 185
7.2 Basic Definitions . . . 188
7.3 The Chain Rule . . . 190
7.4 The Matrix of the Derivative . . . 190
7.5 A Mean Value Inequality . . . 192
7.6 Existence of the Derivative, C^1 Functions . . . 193
7.7 Higher Order Derivatives . . . 196
7.8 Some Standard Notation . . . 197
7.9 The Derivative and the Cartesian Product . . . 198
7.10 Mixed Partial Derivatives . . . 201
7.11 A Cofactor Identity . . . 203
7.12 Newton's Method . . . 205
7.13 Exercises . . . 205

8 Implicit Function Theorem 209
8.1 Statement and Proof of the Theorem . . . 209
8.2 More Derivatives . . . 215
8.3 The Case of $\mathbb{R}^n$ . . . 216
8.4 Exercises . . . 217
8.5 The Method of Lagrange Multipliers . . . 220
8.6 The Taylor Formula . . . 221
8.7 Second Derivative Test . . . 222
8.8 The Rank Theorem . . . 224
8.9 The Local Structure of C^1 Mappings . . . 228
8.10 Invariance of Domain . . . 231
8.11 Exercises . . . 235

II Integration 239

9 Measures and Measurable Functions 241
9.1 Simple Functions and Measurable Functions . . . 241
9.2 Measures and their Properties . . . 246
9.3 Dynkin's Lemma . . . 247
9.4 Measures and Regularity . . . 249
9.5 Outer Measures . . . 253
9.6 Exercises . . . 254
9.7 An Outer Measure on $\mathscr{P}(\mathbb{R})$ . . . 256
9.8 Measures From Outer Measures . . . 259
9.9 Measurable Sets Include Borel Sets? . . . 262
9.10 One Dimensional Lebesgue Stieltjes Measure . . . 263
9.11 Completion of a Measure Space . . . 265
9.12 Vitali Coverings . . . 267

9.13 Differentiation of Increasing Functions . . . 270
9.14 Exercises . . . 272
9.15 Multifunctions and Their Measurability . . . 275
9.15.1 The General Case . . . 275
9.15.2 A Special Case When $\Gamma(\omega)$ Compact . . . 277
9.15.3 Kuratowski's Theorem . . . 278
9.15.4 Measurability of Fixed Points . . . 280
9.15.5 Other Measurability Considerations . . . 280
9.16 Exercises . . . 282

10 The Abstract Lebesgue Integral 283
10.1 Nonnegative Measurable Functions . . . 283
10.1.1 Riemann Integrals For Decreasing Functions . . . 283
10.1.2 The Lebesgue Integral for Nonnegative Functions . . . 284
10.2 Nonnegative Simple Functions . . . 285
10.3 The Monotone Convergence Theorem . . . 287
10.4 Other Definitions . . . 288
10.5 Fatou's Lemma . . . 288
10.6 The Integral's Righteous Algebraic Desires . . . 289
10.7 The Lebesgue Integral, L^1 . . . 289
10.8 The Dominated Convergence Theorem . . . 294
10.9 Some Important General Theory . . . 296
10.9.1 Eggoroff's Theorem . . . 296
10.9.2 The Vitali Convergence Theorem . . . 297
10.10 One Dimensional Lebesgue Stieltjes Integral . . . 300
10.11 The Distribution Function . . . 303
10.12 Good Lambda Inequality . . . 305
10.13 Radon Nikodym Theorem . . . 307
10.14 Abstract Product Measures and Integrals . . . 311
10.15 Exercises . . . 315

11 Regular Measures 319
11.1 Regular Measures in a Metric Space . . . 319
11.2 Differentiation of Radon Measures . . . 321
11.3 Maximal Functions . . . 322
11.4 Strong Estimates for Maximal Function . . . 324
11.5 Symmetric Derivative for Radon Measures . . . 325
11.6 Radon Nikodym Theorem, Radon Measures . . . 327
11.7 Absolutely Continuous Functions . . . 329
11.8 Total Variation . . . 333
11.9 Constructing Measures from Functionals . . . 335
11.10 The p Dimensional Lebesgue Measure . . . 338
11.11 The Brouwer Fixed Point Theorem . . . 341
11.12 Exercises . . . 344
11.13 Change of Variables, Linear Maps . . . 349
11.14 Differentiable Functions and Measurability . . . 352
11.15 Change of Variables, Nonlinear Maps . . . 354
11.16 Mappings which are not One to One . . . 357

11.17 Spherical Coordinates . . . 358
11.18 Exercises . . . 361

12 The L^p Spaces 365
12.1 Basic Inequalities and Properties . . . 365
12.2 Density Considerations . . . 370
12.3 Separability . . . 371
12.4 Continuity of Translation . . . 373
12.5 Mollifiers and Density of Smooth Functions . . . 374
12.6 Smooth Partitions of Unity . . . 377
12.7 Exercises . . . 378

13 Fourier Transforms 383
13.1 Fourier Transforms of Functions in $\mathscr{G}$ . . . 383
13.2 Fourier Transforms of Just about Anything . . . 386
13.2.1 Fourier Transforms of Functions in $L^1(\mathbb{R}^n)$ . . . 389
13.2.2 Fourier Transforms of Functions in $L^2(\mathbb{R}^n)$ . . . 391
13.2.3 The Schwartz Class . . . 394
13.2.4 Convolution . . . 396
13.3 Exercises . . . 398

14 Integration on Manifolds 401
14.1 Relatively Open Sets . . . 401
14.2 Manifolds . . . 401
14.3 The Area Measure on a Manifold . . . 403
14.4 Divergence Theorem . . . 406
14.5 Volumes of Balls in $\mathbb{R}^p$ . . . 410
14.6 Exercises . . . 411

15 Degree Theory 415
15.1 Sard's Lemma and Approximation . . . 417
15.2 Properties of the Degree . . . 424
15.3 Brouwer Fixed Point Theorem . . . 426
15.4 Borsuk's Theorem . . . 428
15.5 Some Applications . . . 430
15.6 Product Formula, Separation Theorem . . . 432
15.7 General Jordan Separation Theorem . . . 437
15.8 Uniqueness of the Degree . . . 441
15.9 Exercises . . . 445

16 Hausdorff Measure 449
16.1 Lipschitz Functions . . . 449
16.2 Lipschitz Functions and Gateaux Derivatives . . . 449
16.3 Rademacher's Theorem . . . 451
16.4 Approximating Lipschitz Maps . . . 453
16.5 Weak Derivatives . . . 455
16.6 Definition of Hausdorff Measures . . . 457
16.7 Properties of Hausdorff Measure . . . 458
16.8 $\mathscr{H}^p$ and m_p . . . 459

16.9 Technical Considerations . . . 461
16.9.1 Steiner Symmetrization . . . 462
16.9.2 The Isodiametric Inequality . . . 463
16.10 The Proper Value of $\beta(p)$ . . . 463
16.11 A Formula for $\alpha(p)$ . . . 464

17 The Area Formula 465
17.1 Estimates for Hausdorff Measure . . . 465
17.2 Comparison Theorems . . . 467
17.3 A Decomposition . . . 468
17.4 Estimates and a Limit . . . 470
17.5 Area Formulas . . . 472
17.6 The Divergence Theorem . . . 476
17.7 The Reynolds Transport Formula . . . 476
17.8 The Coarea Formula . . . 479
17.9 Change of Variables . . . 484
17.10 Integration and the Degree . . . 485

18 Differential Forms 489
18.1 The Wedge Product . . . 493
18.2 The Exterior Derivative . . . 495
18.3 Stokes Theorem . . . 497
18.4 Lipschitz Maps . . . 500
18.5 What Does it Mean? . . . 502
18.6 Examples of $\boldsymbol{r}([\boldsymbol{a},\boldsymbol{b}])$ . . . 505
18.7 Orientation and Degree . . . 505
18.8 Examples of Stoke's Theorem . . . 507
18.8.1 Fundamental Theorem of Calculus . . . 507
18.8.2 Line Integrals for Conservative Fields . . . 507
18.8.3 Green's Theorem . . . 507
18.8.4 Stoke's Theorem from Calculus . . . 508
18.8.5 The Divergence Theorem . . . 510
18.9 Exercises . . . 511

III Abstract Theory 515

19 Hausdorff Spaces and Measures 517
19.1 General Topological Spaces . . . 517
19.2 The Alexander Sub-basis Theorem . . . 523
19.3 The Product Topology and Compactness . . . 524
19.4 Stone Weierstrass Theorem . . . 525
19.4.1 The Case of Locally Compact Sets . . . 525
19.4.2 The Case of Complex Valued Functions . . . 526
19.5 Partitions of Unity . . . 528
19.6 Measures on Hausdorff Spaces . . . 529
19.7 Measures and Positive Linear Functionals . . . 532
19.8 Slicing Measures . . . 537

19.9 Exercises . . . 542

20 Product Measures **545**
20.1 Algebras . . . 545
20.2 Caratheodory Extension Theorem . . . 546
20.3 Kolmogorov Extension Theorem . . . 549
20.4 Exercises . . . 553

21 Banach Spaces **557**
21.1 Theorems Based on Baire Category . . . 557
21.1.1 Baire Category Theorem . . . 557
21.1.2 Uniform Boundedness Theorem . . . 560
21.1.3 Open Mapping Theorem . . . 560
21.1.4 Closed Graph Theorem . . . 562
21.2 Hahn Banach Theorem . . . 564
21.2.1 Partially Ordered Sets . . . 564
21.2.2 Gauge Functions and Hahn Banach Theorem . . . 564
21.2.3 The Complex Version of the Hahn Banach Theorem . . . 566
21.2.4 The Dual Space and Adjoint Operators . . . 567
21.3 Uniform Convexity of L^p . . . 570
21.4 Closed Subspaces . . . 576
21.5 Weak And Weak $*$ Topologies . . . 578
21.5.1 Basic Definitions . . . 578
21.5.2 Banach Alaoglu Theorem . . . 580
21.5.3 Eberlein Smulian Theorem . . . 582
21.6 Differential Equations . . . 585
21.7 Lyapunov Schmidt Procedure . . . 587
21.8 Exercises . . . 591

22 Hilbert Spaces **597**
22.1 Basic Theory . . . 597
22.2 The Hilbert Space $L(U)$ . . . 602
22.3 Approximations in Hilbert Space . . . 604
22.4 Orthonormal Sets . . . 606
22.5 Compact Operators in Hilbert Space . . . 608
22.5.1 Nuclear Operators . . . 612
22.5.2 Hilbert Schmidt Operators . . . 614
22.6 Roots of Positive Linear Maps . . . 617
22.6.1 The Product of Positive Self Adjoint Operators . . . 617
22.6.2 Roots of Positive Self Adjoint Operators . . . 618
22.7 Differential Equations in Banach Space . . . 621
22.8 General Theory of Continuous Semigroups . . . 624
22.8.1 Generators of Semigroups . . . 624
22.8.2 Hille Yosida Theorem . . . 628
22.8.3 An Evolution Equation . . . 632
22.8.4 Adjoints for Closed Operators, Hilbert Space . . . 634
22.8.5 Adjoints, Reflexive Banach Space . . . 637
22.9 Exercises . . . 641

23 Representation Theorems **643**
23.1 Radon Nikodym Theorem . . . 643
23.2 Vector Measures . . . 643
23.3 Representation for the Dual Space of L^p . . . 648
23.4 Weak Compactness . . . 653
23.5 The Dual Space of $L^\infty(\Omega)$ . . . 655
23.6 Non σ Finite Case . . . 658
23.7 The Dual Space of $C_0(X)$ . . . 661
23.7.1 Extending Righteous Functionals . . . 663
23.7.2 The Riesz Representation Theorem . . . 664
23.8 Exercises . . . 665

24 The Bochner Integral **669**
24.1 Strong and Weak Measurability . . . 669
24.1.1 Eggoroff's Theorem . . . 676
24.2 The Bochner Integral . . . 678
24.2.1 Definition and Basic Properties . . . 678
24.2.2 Taking a Closed Operator Out of the Integral . . . 682
24.3 Operator Valued Functions . . . 685
24.3.1 Review of Hilbert Schmidt Theorem . . . 687
24.3.2 Measurable Compact Operators . . . 692
24.4 Fubini's Theorem for Bochner Integrals . . . 692
24.5 The Spaces $L^p(\Omega;X)$ . . . 695
24.6 Measurable Representatives . . . 701
24.7 Vector Measures . . . 703
24.8 The Riesz Representation Theorem . . . 707
24.9 An Example of Polish Space . . . 712
24.10 Weakly Convergent Sequences . . . 714
24.11 Some Embedding Theorems . . . 715
24.12 Conditional Expectation in Banach Spaces . . . 728
24.13 Exercises . . . 732

25 Stone's Theorem and Partitions of Unity **733**
25.1 Partitions of Unity and Stone's Theorem . . . 736
25.2 An Extension Theorem, Retracts . . . 738

IV Stochastic Processes and Probability **741**

26 Independence **743**
26.1 Random Variables and Independence . . . 743
26.2 Convergence in Probability . . . 746
26.3 Kolmogorov Extension Theorem . . . 747
26.4 Independent Events and σ Algebras . . . 748
26.5 Banach Space Valued Random Variables . . . 752
26.6 Reduction to Finite Dimensions . . . 755
26.7 0,1 Laws . . . 755
26.8 Strong Law of Large Numbers . . . 758

27 Analytical Considerations **763**
27.1 The Characteristic Function . . . 763
27.2 Conditional Probability . . . 764
27.3 Conditional Expectation, Sub-martingales . . . 768
27.4 Characteristic Functions and Independence . . . 773
27.5 Characteristic Functions for Measures . . . 777
27.6 Independence in Banach Space . . . 779
27.7 Convolution and Sums . . . 781

28 The Normal Distribution **787**
28.1 The Multivariate Normal Distribution . . . 787
28.2 Linear Combinations . . . 790
28.3 Finding Moments . . . 793
28.4 Prokhorov and Levy Theorems . . . 794
28.5 The Central Limit Theorem . . . 804

29 Martingales **809**
29.1 Conditional Expectation . . . 809
29.2 Conditional Expectation and Independence . . . 812
29.3 Discrete Stochastic Processes . . . 814
29.3.1 Upcrossings . . . 816
29.3.2 The Sub-martingale Convergence Theorem . . . 818
29.3.3 Doob Sub-martingale Estimates . . . 821
29.4 Optional Sampling and Stopping Times . . . 823
29.4.1 Optional Sampling for Martingales . . . 826
29.4.2 Optional Sampling Theorem for Sub-Martingales . . . 827
29.5 Reverse Sub-martingale Convergence Theorem . . . 830
29.6 Strong Law of Large Numbers . . . 832

30 Continuous Stochastic Processes **835**
30.1 Fundamental Definitions and Properties . . . 835
30.2 Kolmogorov Čentsov Continuity Theorem . . . 837
30.3 Filtrations . . . 845
30.4 Martingales and Sub-Martingales . . . 853
30.5 Some Maximal Estimates . . . 854

31 Optional Sampling Theorems **859**
31.1 Review of Discreet Stopping Times . . . 859
31.2 Review of Doob Optional Sampling Theorem . . . 861
31.3 Doob Optional Sampling Continuous Case . . . 862
31.3.1 Stopping Times . . . 862
31.3.2 The Optional Sampling Theorem Continuous Case . . . 867
31.4 Maximal Inequalities and Stopping Times . . . 873
31.5 Continuous Sub-martingale Convergence . . . 877
31.6 Hitting This Before That . . . 881
31.7 The Space $\mathscr{M}_T^p(E)$ . . . 884

32 The Quadratic Variation of a Martingale **887**
32.1 How to Recognize a Martingale . . . 887

32.2 Martingales and Total Variation 890
32.3 The Quadratic Variation 892
32.4 The Covariation 899
32.5 The Burkholder Davis Gundy Inequality 902
32.6 Approximation With Step Functions 908

33 Quadratic Variation and Stochastic Integration 911
33.1 The Stieltjes Integral 919
33.2 The Stochastic Integral When $f(s) \in \mathscr{L}_2(U,H)$ 921
33.3 More on Stopping Times 926
33.4 Local Martingales as Integrators 929
33.5 The Stochastic Integral and the Quadratic Variation 931
33.6 The Case of $f \in L^2(\Omega \times [0,T]; \mathscr{L}_2(U,H))$ 932

Copyright © 2018, You are welcome to use this, including copying it for use in classes or referring to it on line but not to publish it for money. I do constantly upgrade it when I find errors or things which could be improved, and I am constantly finding such things.

Preface

This book is on real and abstract analysis. There are four parts. The last part is an introduction to probability and stochastic processes. A course in multi-variable advanced calculus is contained in the first two. There is more there than one would have time to do and there are some advanced topics which are there because it is a convenient setting for them. Originally the first three parts were written for a multi-variable advanced calculus course which changed over time by inclusion of more advanced topics. However, it was easier to keep track of a single file. The emphasis is on finite dimensional spaces although not totally. There are things which are almost never discussed in multi-variable advanced calculus like the fixed point theorems. However, I think the Brouwer fixed point theorem is one of the most important theorems in mathematics and is being neglected along with its corollaries. I give several proofs in the exercises and in the book. There is too much reliance on these theorems without ever considering proofs. That is why there is a chapter on fixed point theorems. In general, I am trying to include all proofs of theorems instead of making the reader chase after them somewhere else or accept them on faith. I object to attempts to make mathematics functionally like a religion where we are asked to believe the words of authority figures. Of course, when you try to include all the proofs, you run the risk of making mistakes, and I certainly make my share, but one should at least try, even though it also results in a longer book.

I am reviewing a few topics from linear algebra mainly to refresh the memory or to read as needed, but I am assuming that people have had a reasonable course in linear algebra. Linear algebra should come before a book like this one.

I sometimes present really important ideas more than once. Sometimes there is a special case in exercises and later the topic is discussed in the text. I think this can be useful in understanding some of these big theorems. Such duplication may not have been deliberate to begin with, but I have chosen to leave it in many cases.

Finite dimensional degree theory is neglected so there is a chapter on this also, presented as a part of analysis. It seems like it is common to neglect to give a careful treatment of the degree in $\mathbb{R}^p$. This is too bad. You end up missing out on fantastic finite dimensional topology like the Jordan separation theorem. I don't know a good proof for this without something like degree theory. Other somewhat unusual items are things like the Besicovitch covering theorem. It seems to me that this is very important and is certainly one of the most profound theorems I have ever seen. Differentiation theory is carried out for general Radon measures using this covering theorem. This is important because these kinds of measures are encountered in probability. Lebesgue measure is a special case. Abstract theory is presented later and includes the standard theorems on representation, Banach spaces, and so forth. Also included is a treatment of the Kolmogorov extension theorem. This major result is being neglected but, if I understand the situation correctly, it is the foundation for modern probability theory. It belongs in a course on analysis. The Bochner integral is also commonly neglected so I have included a careful treatment of this important topic. Some of us do our research in the context of spaces of Bochner integrable functions involving various function spaces.

There is an introduction to probability and stochastic processes at the end. I have included it because I encountered much of it in my old age and thought it was marvelous mathematics. I was not raised on it and this likely shows. However, it may be that someone can benefit from my efforts to understand this material. I have a hard time with it. There is more in my Topics in analysis book, but that is mostly pretty unorganized because I was gathering it from many different sources for our seminar. I am trying to present a more coherent presentation in this book. This is a very big topic and I must pick what I have found

most interesting, likely offending others who know better than I do what is important.

Chapter 1

Review of Some Linear Algebra

This material can be referred to as needed. It is here in order to make the book self contained.

1.1 The Matrix of a Linear Map

Recall the definition of a linear map. First of all, these need to be defined on a linear space and have values in a linear space.

Definition 1.1.1 *Let $T : V \to W$ be a function. Here V and W are linear spaces. Then $T \in \mathscr{L}(V,W)$ and is a linear map means that for α, β scalars and v_1, v_2 vectors,*

$$T(\alpha v_1 + \beta v_2) = \alpha T v_1 + \beta T v_2$$

Also recall from linear algebra that if you have $T \in \mathscr{L}(\mathbb{F}^n, \mathbb{F}^m)$ it can always be understood in terms of a matrix. That is, there exists an $m \times n$ matrix A such that for all $\boldsymbol{x} \in \mathbb{F}^n$,

$$A\boldsymbol{x} = T\boldsymbol{x}$$

Recall that, from the way we multiply matrices,

$$A = \begin{pmatrix} T\boldsymbol{e}_1 & \cdots & T\boldsymbol{e}_n \end{pmatrix}$$

That is, the i^{th} column is just $T\boldsymbol{e}_i$.

1.2 Block Multiplication of Matrices

Consider the following problem

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} E & F \\ G & H \end{pmatrix}.$$

You know how to do this. You get

$$\begin{pmatrix} AE+BG & AF+BH \\ CE+DG & CF+DH \end{pmatrix}.$$

Now what if instead of numbers, the entries, A, B, C, D, E, F, G are matrices of a size such that the multiplications and additions needed in the above formula all make sense. Would the formula be true in this case?

Suppose A is a matrix of the form

$$A = \begin{pmatrix} A_{11} & \cdots & A_{1m} \\ \vdots & \ddots & \vdots \\ A_{r1} & \cdots & A_{rm} \end{pmatrix} \tag{1.1}$$

where A_{ij} is a $s_i \times p_j$ matrix where s_i is constant for $j = 1, \cdots, m$ for each $i = 1, \cdots, r$. Such a matrix is called a **block matrix,** also a **partitioned matrix**. How do you get the block

A_{ij}? Here is how for A an $m \times n$ matrix:

$$\overbrace{\begin{pmatrix} \mathbf{0} & I_{s_i \times s_i} & \mathbf{0} \end{pmatrix}}^{s_i \times m} A \overbrace{\begin{pmatrix} \mathbf{0} \\ I_{p_j \times p_j} \\ \mathbf{0} \end{pmatrix}}^{n \times p_j}. \tag{1.2}$$

In the block column matrix on the right, you need to have $c_j - 1$ rows of zeros above the small $p_j \times p_j$ identity matrix where the columns of A involved in A_{ij} are $c_j, \cdots, c_j + p_j - 1$ and in the block row matrix on the left, you need to have $r_i - 1$ columns of zeros to the left of the $s_i \times s_i$ identity matrix where the rows of A involved in A_{ij} are $r_i, \cdots, r_i + s_i$. An important observation to make is that the matrix on the right specifies columns to use in the block and the one on the left specifies the rows. Thus the block A_{ij}, in this case, is a matrix of size $s_i \times p_j$. There is no overlap between the blocks of A. Thus the identity $n \times n$ identity matrix corresponding to multiplication on the right of A is of the form

$$\begin{pmatrix} I_{p_1 \times p_1} & & 0 \\ & \ddots & \\ 0 & & I_{p_m \times p_m} \end{pmatrix},$$

where these little identity matrices don't overlap. A similar conclusion follows from consideration of the matrices $I_{s_i \times s_i}$. Note that in (1.2), the matrix on the right is a block column matrix for the above block diagonal matrix, and the matrix on the left in (1.2) is a block row matrix taken from a similar block diagonal matrix consisting of the $I_{s_i \times s_i}$.

Next consider the question of multiplication of two block matrices. Let B be a block matrix of the form

$$\begin{pmatrix} B_{11} & \cdots & B_{1p} \\ \vdots & \ddots & \vdots \\ B_{r1} & \cdots & B_{rp} \end{pmatrix} \tag{1.3}$$

and A is a block matrix of the form

$$\begin{pmatrix} A_{11} & \cdots & A_{1m} \\ \vdots & \ddots & \vdots \\ A_{p1} & \cdots & A_{pm} \end{pmatrix} \tag{1.4}$$

such that for all i, j, it makes sense to multiply $B_{is}A_{sj}$ for all $s \in \{1, \cdots, p\}$. (That is the two matrices B_{is} and A_{sj} are conformable.) and that for fixed ij, it follows that $B_{is}A_{sj}$ is the same size for each s so that it makes sense to write $\sum_s B_{is}A_{sj}$.

The following theorem says essentially that when you take the product of two matrices, you can partition both matrices, formally multiply the blocks to get another block matrix and this one will be BA partitioned. Before presenting this theorem, here is a simple lemma which is really a special case of the theorem.

Lemma 1.2.1 *Consider the following product.*

$$\begin{pmatrix} \mathbf{0} \\ I \\ \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{0} & I & \mathbf{0} \end{pmatrix}$$

where the first is $n \times r$ and the second is $r \times n$. The small identity matrix I is an $r \times r$ matrix and there are l zero rows above I and l zero columns to the left of I in the right matrix. Then the product of these matrices is a block matrix of the form

$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Proof: From the definition of matrix multiplication, the product is

$$\begin{pmatrix} \begin{pmatrix} 0 \\ I \\ 0 \end{pmatrix} 0 & \cdots & \begin{pmatrix} 0 \\ I \\ 0 \end{pmatrix} e_1 & \cdots & \begin{pmatrix} 0 \\ I \\ 0 \end{pmatrix} e_r & \cdots & \begin{pmatrix} 0 \\ I \\ 0 \end{pmatrix} 0 \end{pmatrix}$$

which yields the claimed result. In the formula e_j refers to the column vector of length r which has a 1 in the j^{th} position. This proves the lemma. ■

Theorem 1.2.2 *Let B be a $q \times p$ block matrix as in (1.3) and let A be a $p \times n$ block matrix as in (1.4) such that B_{is} is conformable with A_{sj} and each product, $B_{is}A_{sj}$ for $s = 1, \cdots, p$ is of the same size, so that they can be added. Then BA can be obtained as a block matrix such that the ij^{th} block is of the form*

$$\sum_s B_{is}A_{sj}. \tag{1.5}$$

Proof: From (1.2)

$$B_{is}A_{sj} = \begin{pmatrix} 0 & I_{r_i \times r_i} & 0 \end{pmatrix} B \begin{pmatrix} 0 \\ I_{p_s \times p_s} \\ 0 \end{pmatrix} \begin{pmatrix} 0 & I_{p_s \times p_s} & 0 \end{pmatrix} A \begin{pmatrix} 0 \\ I_{q_j \times q_j} \\ 0 \end{pmatrix}$$

where here it is assumed B_{is} is $r_i \times p_s$ and A_{sj} is $p_s \times q_j$. The product involves the s^{th} block in the i^{th} row of blocks for B and the s^{th} block in the j^{th} column of A. Thus there are the same number of rows above the $I_{p_s \times p_s}$ as there are columns to the left of $I_{p_s \times p_s}$ in those two inside matrices. Then from Lemma 1.2.1

$$\begin{pmatrix} 0 \\ I_{p_s \times p_s} \\ 0 \end{pmatrix} \begin{pmatrix} 0 & I_{p_s \times p_s} & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & I_{p_s \times p_s} & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Since the blocks of small identity matrices do not overlap,

$$\sum_s \begin{pmatrix} 0 & 0 & 0 \\ 0 & I_{p_s \times p_s} & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} I_{p_1 \times p_1} & & 0 \\ & \ddots & \\ 0 & & I_{p_p \times p_p} \end{pmatrix} = I,$$

and so,

$$\sum_s B_{is}A_{sj} = \sum_s \begin{pmatrix} 0 & I_{r_i \times r_i} & 0 \end{pmatrix} B \begin{pmatrix} 0 \\ I_{p_s \times p_s} \\ 0 \end{pmatrix} \begin{pmatrix} 0 & I_{p_s \times p_s} & 0 \end{pmatrix} A \begin{pmatrix} 0 \\ I_{q_j \times q_j} \\ 0 \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{0} & I_{r_i \times r_i} & \mathbf{0} \end{pmatrix} B \sum_s \begin{pmatrix} \mathbf{0} \\ I_{p_s \times p_s} \\ \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{0} & I_{p_s \times p_s} & \mathbf{0} \end{pmatrix} A \begin{pmatrix} \mathbf{0} \\ I_{q_j \times q_j} \\ \mathbf{0} \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{0} & I_{r_i \times r_i} & \mathbf{0} \end{pmatrix} BIA \begin{pmatrix} \mathbf{0} \\ I_{q_j \times q_j} \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{0} & I_{r_i \times r_i} & \mathbf{0} \end{pmatrix} BA \begin{pmatrix} \mathbf{0} \\ I_{q_j \times q_j} \\ \mathbf{0} \end{pmatrix}$$

which equals the ij^{th} block of BA. Hence the ij^{th} block of BA equals the formal multiplication according to matrix multiplication,

$$\sum_s B_{is} A_{sj}.$$

This proves the theorem. ■

Example 1.2.3 *Multiply the following pair of partitioned matrices using the above theorem by multiplying the blocks as described above and then in the conventional manner.*

$$\left(\begin{array}{cc|c} 1 & 2 & 3 \\ \hline -1 & 2 & 3 \\ 3 & -2 & 1 \end{array}\right) \left(\begin{array}{c|cc} 1 & -1 & 2 \\ 2 & 3 & 0 \\ \hline -2 & 2 & 1 \end{array}\right)$$

Doing it in terms of the blocks, this yields, after the indicated multiplications of the blocks,

$$\left(\begin{array}{c|c} 5+(-6) & \begin{pmatrix} 5 & 2 \end{pmatrix} + 3 \begin{pmatrix} 2 & 1 \end{pmatrix} \\ \hline \begin{pmatrix} 3 \\ -1 \end{pmatrix} + \begin{pmatrix} 3 \\ 1 \end{pmatrix} (-2) & \begin{pmatrix} 7 & -2 \\ -9 & 6 \end{pmatrix} + \begin{pmatrix} 6 & 3 \\ 2 & 1 \end{pmatrix} \end{array}\right)$$

This is

$$\left(\begin{array}{c|c} -1 & \begin{pmatrix} 11 & 5 \end{pmatrix} \\ \hline \begin{pmatrix} -3 \\ -3 \end{pmatrix} & \begin{pmatrix} 13 & 1 \\ -7 & 7 \end{pmatrix} \end{array}\right)$$

Multiplying it out the usual way, you have

$$\begin{pmatrix} 1 & 2 & 3 \\ -1 & 2 & 3 \\ 3 & -2 & 1 \end{pmatrix} \begin{pmatrix} 1 & -1 & 2 \\ 2 & 3 & 0 \\ -2 & 2 & 1 \end{pmatrix} = \begin{pmatrix} -1 & 11 & 5 \\ -3 & 13 & 1 \\ -3 & -7 & 7 \end{pmatrix}$$

you see this is the same thing without the partition lines.

1.3 Schur's Theorem

For some reason, not understood by me, Schur's theorem is often neglected in beginning linear algebra. This is too bad because it is one of the best theorems in linear algebra. Here $|\cdot|$ denotes the usual norm in $\mathbb{C}^n$ given by

$$|\boldsymbol{x}|^2 \equiv \sum_{j=1}^{n} |x_j|^2$$

Definition 1.3.1 *A complex $n \times n$ matrix U is said to be unitary if $U^*U = I$. Here U^* is the transpose of the conjugate of U. The matrix is unitary if and only if its columns form an orthonormal set in $\mathbb{C}^n$. This follows from the way we multiply matrices in which the ij^{th} entry of U^*U is obtained by taking the conjugate of the i^{th} row of U times the j^{th} column of U.*

Theorem 1.3.2 *(Schur) Let A be a complex $n \times n$ matrix. Then there exists a unitary matrix U such that*

$$U^*AU = T, \tag{1.6}$$

where T is an upper triangular matrix having the eigenvalues of A on the main diagonal, listed with multiplicity[1].

Proof: The theorem is clearly true if A is a 1×1 matrix. Just let $U = 1$, the 1×1 matrix which has entry 1. Suppose it is true for $(n-1) \times (n-1)$ matrices and let A be an $n \times n$ matrix. Then let $\boldsymbol{v}_1$ be a unit eigenvector for A. Then there exists λ_1 such that

$$A\boldsymbol{v}_1 = \lambda_1 \boldsymbol{v}_1, \ |\boldsymbol{v}_1| = 1.$$

Extend $\{\boldsymbol{v}_1\}$ to a basis and then use the Gram - Schmidt process to obtain

$$\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n\}$$

an orthonormal basis of $\mathbb{C}^n$. Let U_0 be a matrix whose i^{th} column is $\boldsymbol{v}_i$. Then from the definition of a unitary matrix Definition 1.3.1, it follows that U_0 is unitary. Consider $U_0^*AU_0$.

$$U_0^*AU_0 = \begin{pmatrix} \boldsymbol{v}_1^* \\ \vdots \\ \boldsymbol{v}_n^* \end{pmatrix} \begin{pmatrix} A\boldsymbol{v}_1 & \cdots & A\boldsymbol{v}_n \end{pmatrix} = \begin{pmatrix} \boldsymbol{v}_1^* \\ \vdots \\ \boldsymbol{v}_n^* \end{pmatrix} \begin{pmatrix} \lambda_1 \boldsymbol{v}_1 & \cdots & A\boldsymbol{v}_n \end{pmatrix}$$

Thus $U_0^*AU_0$ is of the form

$$\begin{pmatrix} \lambda_1 & \boldsymbol{a} \\ \boldsymbol{0} & A_1 \end{pmatrix}$$

where A_1 is an $n-1 \times n-1$ matrix. Now by induction, there exists an $(n-1) \times (n-1)$ unitary matrix $\widetilde{U}_1$ such that

$$\widetilde{U}_1^* A_1 \widetilde{U}_1 = T_{n-1},$$

an upper triangular matrix. Consider

$$U_1 \equiv \begin{pmatrix} 1 & \boldsymbol{0} \\ \boldsymbol{0} & \widetilde{U}_1 \end{pmatrix}.$$

An application of block multiplication shows that U_1 is a unitary matrix and also that

$$U_1^* U_0^* A U_0 U_1 = \begin{pmatrix} 1 & \boldsymbol{0} \\ \boldsymbol{0} & \widetilde{U}_1^* \end{pmatrix} \begin{pmatrix} \lambda_1 & * \\ \boldsymbol{0} & A_1 \end{pmatrix} \begin{pmatrix} 1 & \boldsymbol{0} \\ \boldsymbol{0} & \widetilde{U}_1 \end{pmatrix} = \begin{pmatrix} \lambda_1 & * \\ \boldsymbol{0} & T_{n-1} \end{pmatrix} = T$$

where T is upper triangular. Then let $U = U_0U_1$. Since $(U_0U_1)^* = U_1^*U_0^*$, it follows that A is similar to T and that U_0U_1 is unitary. Hence A and T have the same characteristic

[1] 'Listed with multiplicity' means that the diagonal entries are repeated according to their multiplicity as roots of the characteristic equation.

polynomials, and since the eigenvalues of T are the diagonal entries listed with multiplicity, this proves the theorem. ■

The same argument yields the following corollary in the case where A has real entries. The only difference is the use of the real inner product instead of the complex inner product.

Corollary 1.3.3 *Let A be a real $n \times n$ matrix which has only real eigenvalues. Then there exists a real orthogonal matrix Q such that*

$$Q^T A Q = T$$

where T is an upper triangular matrix having the eigenvalues of A on the main diagonal, listed with multiplicity.

Proof: This follows by observing that if all eigenvalues are real, then corresponding to each real eigenvalue, there exists a real eigenvector. Thus the argument of the above theorem applies with the real inner product in $\mathbb{R}^n$. ■

1.4 Hermitian and Symmetric Matrices

A complex $n \times n$ matrix A with $A^* = A$ is said to be **Hermitian**. A real $n \times n$ matrix A with $A^T = A$ is said to be **symmetric**. In either case, note that for $\langle \cdot, \cdot \rangle$ the inner product in $\mathbb{C}^n$,

$$\langle A\boldsymbol{u}, \boldsymbol{v} \rangle = (A\boldsymbol{u})^T \bar{\boldsymbol{v}} = \boldsymbol{u}^T A^T \bar{\boldsymbol{v}} = \boldsymbol{u}^T \overline{A} \bar{\boldsymbol{v}} = \langle \boldsymbol{u}, A\boldsymbol{v} \rangle.$$

Thus, as a numerical example, the matrix

$$\begin{pmatrix} 1 & 1-i \\ 1+i & 2 \end{pmatrix}$$

is Hermitian, while

$$\begin{pmatrix} 1 & -1 & -2 \\ -1 & 2 & 4 \\ -2 & 4 & 3 \end{pmatrix}$$

is symmetric. Hermitian matrices are named in honor of the French mathematician Charles Hermite (1822–1901).

With Schur's theorem, the theorem on diagonalization of a Hermitian matrix follows.

Theorem 1.4.1 *Let A be Hermitian. Then the eigenvalues of A are all real, and there exists a unitary matrix U such that*

$$U^* A U = D,$$

a diagonal matrix whose diagonal entries are the eigenvalues of A listed with multiplicity. In case A is symmetric, U may be taken to be an orthogonal matrix. The columns of U form an orthonormal basis of eigenvectors of A.

Proof: By Schur's theorem and the assumption that A is Hermitian, there exists a triangular matrix T, whose diagonal entries are the eigenvalues of A listed with multiplicity, and a unitary matrix U such that

$$T = U^* A U = U^* A^* U = (U^* A U)^* = T^*.$$

It follows from this that T is a diagonal matrix and has all real entries down the main diagonal. Hence the eigenvalues of A are real. If A is symmetric (real and Hermitian) it follows from Corollary 1.3.3 that U may be taken to be orthogonal (The columns are an orthonormal set in the inner product of $\mathbb{R}^n$).

That the columns of U form an orthonormal basis of eigenvectors of A, follows right away from the definition of matrix multiplication which implies that if $\boldsymbol{u}_i$ is a column of U, then $A\boldsymbol{u}_i = \text{column } i \text{ of } (UD) = \lambda_i \boldsymbol{u}_i$. ■

1.5 The Right Polar Factorization

The right polar factorization involves writing a matrix as a product of two other matrices, one which preserves distances and the other which stretches and distorts. First here are some lemmas which review and add to many of the topics discussed so far about adjoints and orthonormal sets and such things. This is of fundamental significance in geometric measure theory and also in continuum mechanics. Not surprisingly the stress should depend on the part which stretches and distorts. See [23].

Lemma 1.5.1 *Let A be a Hermitian matrix such that all its eigenvalues are nonnegative. Then there exists a Hermitian matrix $A^{1/2}$ such that $A^{1/2}$ has all nonnegative eigenvalues and $\left(A^{1/2}\right)^2 = A$.*

Proof: Since A is Hermitian, there exists a diagonal matrix D having all real nonnegative entries and a unitary matrix U such that $A = U^*DU$. This is from Theorem 1.4.1 above. Then denote by $D^{1/2}$ the matrix which is obtained by replacing each diagonal entry of D with its square root. Thus $D^{1/2}D^{1/2} = D$. Then define

$$A^{1/2} \equiv U^*D^{1/2}U.$$

Then

$$\left(A^{1/2}\right)^2 = U^*D^{1/2}UU^*D^{1/2}U = U^*DU = A.$$

Since $D^{1/2}$ is real,

$$\left(U^*D^{1/2}U\right)^* = U^*\left(D^{1/2}\right)^*(U^*)^* = U^*D^{1/2}U$$

so $A^{1/2}$ is Hermitian. ■

Next it is helpful to recall the Gram Schmidt algorithm and observe a certain property stated in the next lemma.

Lemma 1.5.2 *Suppose $\{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_r, \boldsymbol{v}_{r+1}, \cdots, \boldsymbol{v}_p\}$ is a linearly independent set of vectors such that $\{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_r\}$ is an orthonormal set of vectors. Then when the Gram Schmidt process is applied to the vectors in the given order, it will not change any of the $\boldsymbol{w}_1, \cdots, \boldsymbol{w}_r$.*

Proof: Let $\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_p\}$ be the orthonormal set delivered by the Gram Schmidt process. Then $\boldsymbol{u}_1 = \boldsymbol{w}_1$ because by definition, $\boldsymbol{u}_1 \equiv \boldsymbol{w}_1/|\boldsymbol{w}_1| = \boldsymbol{w}_1$. Now suppose $\boldsymbol{u}_j = \boldsymbol{w}_j$ for all $j \leq k \leq r$. Then if $k < r$, consider the definition of $\boldsymbol{u}_{k+1}$.

$$\boldsymbol{u}_{k+1} \equiv \frac{\boldsymbol{w}_{k+1} - \sum_{j=1}^{k+1} (\boldsymbol{w}_{k+1}, \boldsymbol{u}_j)\, \boldsymbol{u}_j}{\left|\boldsymbol{w}_{k+1} - \sum_{j=1}^{k+1} (\boldsymbol{w}_{k+1}, \boldsymbol{u}_j)\, \boldsymbol{u}_j\right|}$$

By induction, $\boldsymbol{u}_j = \boldsymbol{w}_j$ and so this reduces to $\boldsymbol{w}_{k+1}/|\boldsymbol{w}_{k+1}| = \boldsymbol{w}_{k+1}$. ■

This lemma immediately implies the following lemma.

Lemma 1.5.3 *Let V be a subspace of dimension p and let $\{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_r\}$ be an orthonormal set of vectors in V. Then this orthonormal set of vectors may be extended to an orthonormal basis for V,*

$$\{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_r, \boldsymbol{y}_{r+1}, \cdots, \boldsymbol{y}_p\}$$

Proof: First extend the given linearly independent set $\{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_r\}$ to a basis for V and then apply the Gram Schmidt theorem to the resulting basis. Since $\{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_r\}$ is orthonormal it follows from Lemma 1.5.2 the result is of the desired form, an orthonormal basis extending $\{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_r\}$. ■

Here is another lemma about preserving distance.

Lemma 1.5.4 *Suppose R is an $m \times n$ matrix with $m \geq n$ and R preserves distances. Then $R^*R = I$. Also, if R takes an orthonormal basis to an orthonormal set, then R must preserve distances.*

Proof: Since R preserves distances, $|R\boldsymbol{x}| = |\boldsymbol{x}|$ for every $\boldsymbol{x}$. Therefore from the axioms of the dot product,

$$\begin{aligned} & |\boldsymbol{x}|^2 + |\boldsymbol{y}|^2 + (\boldsymbol{x},\boldsymbol{y}) + (\boldsymbol{y},\boldsymbol{x}) = |\boldsymbol{x}+\boldsymbol{y}|^2 = (R(\boldsymbol{x}+\boldsymbol{y}), R(\boldsymbol{x}+\boldsymbol{y})) \\ = \; & (R\boldsymbol{x}, R\boldsymbol{x}) + (R\boldsymbol{y}, R\boldsymbol{y}) + (R\boldsymbol{x}, R\boldsymbol{y}) + (R\boldsymbol{y}, R\boldsymbol{x}) \\ = \; & |\boldsymbol{x}|^2 + |\boldsymbol{y}|^2 + (R^*R\boldsymbol{x}, \boldsymbol{y}) + (\boldsymbol{y}, R^*R\boldsymbol{x}) \end{aligned}$$

and so for all $\boldsymbol{x}, \boldsymbol{y}$,

$$(R^*R\boldsymbol{x} - \boldsymbol{x}, \boldsymbol{y}) + (\boldsymbol{y}, R^*R\boldsymbol{x} - \boldsymbol{x}) = 0$$

Hence for all $\boldsymbol{x}, \boldsymbol{y}$,

$$\mathrm{Re}\,(R^*R\boldsymbol{x} - \boldsymbol{x}, \boldsymbol{y}) = 0$$

Now for a $\boldsymbol{x}, \boldsymbol{y}$ given, choose $\alpha \in \mathbb{C}$ such that

$$\alpha\,(R^*R\boldsymbol{x} - \boldsymbol{x}, \boldsymbol{y}) = |(R^*R\boldsymbol{x} - \boldsymbol{x}, \boldsymbol{y})|$$

Then

$$0 = \mathrm{Re}\,(R^*R\boldsymbol{x} - \boldsymbol{x}, \overline{\alpha}\boldsymbol{y}) = \mathrm{Re}\,\alpha\,(R^*R\boldsymbol{x} - \boldsymbol{x}, \boldsymbol{y}) = |(R^*R\boldsymbol{x} - \boldsymbol{x}, \boldsymbol{y})|$$

Thus $|(R^*R\boldsymbol{x} - \boldsymbol{x}, \boldsymbol{y})| = 0$ for all $\boldsymbol{x}, \boldsymbol{y}$ because the given $\boldsymbol{x}, \boldsymbol{y}$ were arbitrary. Let $\boldsymbol{y} = R^*R\boldsymbol{x} - \boldsymbol{x}$ to conclude that for all $\boldsymbol{x}$,

$$R^*R\boldsymbol{x} - \boldsymbol{x} = \boldsymbol{0}$$

which says $R^*R = I$ since $\boldsymbol{x}$ is arbitrary.

Consider the last claim. Let $R : \mathbb{F}^n \to \mathbb{F}^m$ such that $\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_n\}$ is an orthonormal basis for $\mathbb{F}^n$ and $\{R\boldsymbol{u}_1, \cdots, R\boldsymbol{u}_n\}$ is also an orthormal set, then

$$\left| R\left(\sum_i x_i \boldsymbol{u}_i\right) \right|^2 = \left| \sum_i x_i R\boldsymbol{u}_i \right|^2 = \sum_i |x_i|^2 = \left| \sum_i x_i \boldsymbol{u}_i \right|^2 \quad ■$$

With this preparation, here is the big theorem about the right polar factorization.

Theorem 1.5.5 *Let F be an $m \times n$ matrix where $m \geq n$. Then there exists a Hermitian $n \times n$ matrix U which has all nonnegative eigenvalues and an $m \times n$ matrix R which satisfies $R^*R = I$ such that $F = RU$.*

Proof: Consider F^*F. This is a Hermitian matrix because

$$(F^*F)^* = F^*(F^*)^* = F^*F$$

Also the eigenvalues of the $n \times n$ matrix F^*F are all nonnegative. This is because if $\boldsymbol{x}$ is an eigenvalue,

$$\lambda(\boldsymbol{x},\boldsymbol{x}) = (F^*F\boldsymbol{x},\boldsymbol{x}) = (F\boldsymbol{x},F\boldsymbol{x}) \geq 0.$$

Therefore, by Lemma 1.5.1, there exists an $n \times n$ Hermitian matrix U having all nonnegative eigenvalues such that

$$U^2 = F^*F.$$

Consider the subspace $U(\mathbb{F}^n)$. Let $\{U\boldsymbol{x}_1, \cdots, U\boldsymbol{x}_r\}$ be an orthonormal basis for

$$U(\mathbb{F}^n) \subseteq \mathbb{F}^n.$$

Note that $U(\mathbb{F}^n)$ might not be all of $\mathbb{F}^n$. Using Lemma 1.5.3, extend to an orthonormal basis for all of $\mathbb{F}^n$,

$$\{U\boldsymbol{x}_1, \cdots, U\boldsymbol{x}_r, \boldsymbol{y}_{r+1}, \cdots, \boldsymbol{y}_n\}.$$

Next observe that $\{F\boldsymbol{x}_1, \cdots, F\boldsymbol{x}_r\}$ is also an orthonormal set of vectors in $\mathbb{F}^m$. This is because

$$\begin{aligned}(F\boldsymbol{x}_k, F\boldsymbol{x}_j) &= (F^*F\boldsymbol{x}_k, \boldsymbol{x}_j) = \left(U^2\boldsymbol{x}_k, \boldsymbol{x}_j\right) \\ &= (U\boldsymbol{x}_k, U^*\boldsymbol{x}_j) = (U\boldsymbol{x}_k, U\boldsymbol{x}_j) = \delta_{jk}\end{aligned}$$

Therefore, from Lemma 1.5.3 again, this orthonormal set of vectors can be extended to an orthonormal basis for $\mathbb{F}^m$,

$$\{F\boldsymbol{x}_1, \cdots, F\boldsymbol{x}_r, \boldsymbol{z}_{r+1}, \cdots, \boldsymbol{z}_m\}$$

Thus there are at least as many $\boldsymbol{z}_k$ as there are $\boldsymbol{y}_j$. Now for $\boldsymbol{x} \in \mathbb{F}^n$, since

$$\{U\boldsymbol{x}_1, \cdots, U\boldsymbol{x}_r, \boldsymbol{y}_{r+1}, \cdots, \boldsymbol{y}_n\}$$

is an orthonormal basis for $\mathbb{F}^n$, there exist unique scalars,

$$c_1 \cdots, c_r, d_{r+1}, \cdots, d_n$$

such that

$$\boldsymbol{x} = \sum_{k=1}^{r} c_k U\boldsymbol{x}_k + \sum_{k=r+1}^{n} d_k \boldsymbol{y}_k$$

Define

$$R\boldsymbol{x} \equiv \sum_{k=1}^{r} c_k F\boldsymbol{x}_k + \sum_{k=r+1}^{n} d_k \boldsymbol{z}_k \tag{1.7}$$

Then also there exist scalars b_k such that

$$U\boldsymbol{x} = \sum_{k=1}^{r} b_k U\boldsymbol{x}_k$$

and so from 1.7,

$$RU\,\boldsymbol{x} = \sum_{k=1}^{r} b_k F\boldsymbol{x}_k = F\left(\sum_{k=1}^{r} b_k \boldsymbol{x}_k\right)$$

Is $F\left(\sum_{k=1}^{r} b_k \boldsymbol{x}_k\right) = F(\boldsymbol{x})$?

$$\left(F\left(\sum_{k=1}^{r} b_k \boldsymbol{x}_k\right) - F(\boldsymbol{x}), F\left(\sum_{k=1}^{r} b_k \boldsymbol{x}_k\right) - F(\boldsymbol{x})\right)$$

$$\begin{aligned}
&= \left((F^*F)\left(\sum_{k=1}^{r} b_k \boldsymbol{x}_k - \boldsymbol{x}\right), \left(\sum_{k=1}^{r} b_k \boldsymbol{x}_k - \boldsymbol{x}\right)\right) \\
&= \left(U^2\left(\sum_{k=1}^{r} b_k \boldsymbol{x}_k - \boldsymbol{x}\right), \left(\sum_{k=1}^{r} b_k \boldsymbol{x}_k - \boldsymbol{x}\right)\right) \\
&= \left(U\left(\sum_{k=1}^{r} b_k \boldsymbol{x}_k - \boldsymbol{x}\right), U\left(\sum_{k=1}^{r} b_k \boldsymbol{x}_k - \boldsymbol{x}\right)\right) \\
&= \left(\sum_{k=1}^{r} b_k U\boldsymbol{x}_k - U\boldsymbol{x}, \sum_{k=1}^{r} b_k U\boldsymbol{x}_k - U\boldsymbol{x}\right) = 0
\end{aligned}$$

Therefore, $F\left(\sum_{k=1}^{r} b_k \boldsymbol{x}_k\right) = F(\boldsymbol{x})$ and this shows $RU\boldsymbol{x} = F\boldsymbol{x}$**.** From 1.7 it follows that R maps an orthonormal set to an orthonormal set and so R preserves distances. Therefore, by Lemma 1.5.4 $R^*R = I$. ■

1.6 Elementary matrices

The elementary matrices result from doing a row operation to the identity matrix.

As before, everything will apply to matrices having coefficients in an arbitrary field of scalars, although we will mainly feature the real numbers in the examples.

Definition 1.6.1 *The row operations consist of the following*

1. *Switch two rows.*
2. *Multiply a row by a nonzero number.*
3. *Replace a row by the same row added to a multiple of another row.*

We refer to these as the row operations of type 1,2, and 3 respectively.

The elementary matrices are given in the following definition.

Definition 1.6.2 *The elementary matrices consist of those matrices which result by applying a row operation to an identity matrix. Those which involve switching rows of the identity are called permutation matrices. More generally, a permutation matrix is a matrix which comes by permuting the rows of the identity matrix, not just switching two rows.*

As an example of why these elementary matrices are interesting, consider the following.

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a & b & c & d \\ x & y & z & w \\ f & g & h & i \end{pmatrix} = \begin{pmatrix} x & y & z & w \\ a & b & c & d \\ f & g & h & i \end{pmatrix}.$$

A 3×4 matrix was multiplied on the left by an elementary matrix which was obtained from row operation 1 applied to the identity matrix. This resulted in applying the operation 1 to the given matrix. This is what happens in general.

Now consider what these elementary matrices look like. First P_{ij}, which involves switching row i and row j of the identity where $i < j$. We write

$$I = \begin{pmatrix} \boldsymbol{r}_1 \\ \vdots \\ \boldsymbol{r}_i \\ \vdots \\ \boldsymbol{r}_j \\ \vdots \\ \boldsymbol{r}_n \end{pmatrix}$$

where

$$\boldsymbol{r}_j = (0 \cdots 1 \cdots 0)$$

with the 1 in the j^{th} position from the left.

This matrix P^{ij} is of the form

$$\begin{pmatrix} \boldsymbol{r}_1 \\ \vdots \\ \boldsymbol{r}_j \\ \vdots \\ \boldsymbol{r}_i \\ \vdots \\ \boldsymbol{r}_n \end{pmatrix}.$$

Now consider what this does to a column vector.

$$\begin{pmatrix} \boldsymbol{r}_1 \\ \vdots \\ \boldsymbol{r}_j \\ \vdots \\ \boldsymbol{r}_i \\ \vdots \\ \boldsymbol{r}_n \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_i \\ \vdots \\ v_j \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} v_1 \\ \vdots \\ v_j \\ \vdots \\ v_i \\ \vdots \\ v_n \end{pmatrix}.$$

Now we try multiplication of a matrix on the left by this elementary matrix P^{ij}. Consider

$$P^{ij}\begin{pmatrix} a_{11} & a_{12} & \cdots & \cdots & \cdots & \cdots & a_{1p} \\ \vdots & \vdots & & & & & \vdots \\ a_{i1} & a_{i2} & \cdots & \cdots & \cdots & \cdots & a_{ip} \\ \vdots & \vdots & & & & & \vdots \\ a_{j1} & a_{j2} & \cdots & \cdots & \cdots & \cdots & a_{jp} \\ \vdots & \vdots & & & & & \vdots \\ a_{n1} & a_{n2} & \cdots & \cdots & \cdots & \cdots & a_{np} \end{pmatrix}.$$

From the way you multiply matrices this is a matrix which has the indicated columns.

$$\left(P^{ij}\begin{pmatrix} a_{11} \\ \vdots \\ a_{i1} \\ \vdots \\ a_{j1} \\ \vdots \\ a_{n1} \end{pmatrix}, P^{ij}\begin{pmatrix} a_{12} \\ \vdots \\ a_{i2} \\ \vdots \\ a_{j2} \\ \vdots \\ a_{n2} \end{pmatrix}, \cdots, P^{ij}\begin{pmatrix} a_{1p} \\ \vdots \\ a_{ip} \\ \vdots \\ a_{jp} \\ \vdots \\ a_{np} \end{pmatrix}\right)$$

$$= \left(\begin{pmatrix} a_{11} \\ \vdots \\ a_{j1} \\ \vdots \\ a_{i1} \\ \vdots \\ a_{n1} \end{pmatrix}, \begin{pmatrix} a_{12} \\ \vdots \\ a_{j2} \\ \vdots \\ a_{i2} \\ \vdots \\ a_{n2} \end{pmatrix}, \cdots, \begin{pmatrix} a_{1p} \\ \vdots \\ a_{jp} \\ \vdots \\ a_{ip} \\ \vdots \\ a_{np} \end{pmatrix}\right)$$

$$= \begin{pmatrix} a_{11} & a_{12} & \cdots & \cdots & \cdots & \cdots & a_{1p} \\ \vdots & \vdots & & & & & \vdots \\ a_{j1} & a_{j2} & \cdots & \cdots & \cdots & \cdots & a_{jp} \\ \vdots & \vdots & & & & & \vdots \\ a_{i1} & a_{i2} & \cdots & \cdots & \cdots & \cdots & a_{ip} \\ \vdots & \vdots & & & & & \vdots \\ a_{n1} & a_{n2} & \cdots & \cdots & \cdots & \cdots & a_{np} \end{pmatrix}.$$

This has established the following lemma.

Lemma 1.6.3 *Let P^{ij} denote the elementary matrix which involves switching the i^{th} and the j^{th} rows of I. Then if P^{ij}, A are conformable, we have*

$$P^{ij}A = B$$

where B is obtained from A by switching the i^{th} and the j^{th} rows.

Next consider the row operation which involves multiplying the i^{th} row by a nonzero constant, c. We write

$$I = \begin{pmatrix} \boldsymbol{r}_1 \\ \boldsymbol{r}_2 \\ \vdots \\ \boldsymbol{r}_n \end{pmatrix}$$

where

$$\boldsymbol{r}_j = (0 \cdots 1 \cdots 0)$$

with the 1 in the j^{th} position from the left. The elementary matrix which results from applying this operation to the i^{th} row of the identity matrix is of the form

$$E(c,i) = \begin{pmatrix} \boldsymbol{r}_1 \\ \vdots \\ c\boldsymbol{r}_i \\ \vdots \\ \boldsymbol{r}_n \end{pmatrix}.$$

Now consider what this does to a column vector.

$$\begin{pmatrix} \boldsymbol{r}_1 \\ \vdots \\ c\boldsymbol{r}_i \\ \vdots \\ \boldsymbol{r}_n \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_i \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} v_1 \\ \vdots \\ cv_i \\ \vdots \\ v_n \end{pmatrix}.$$

Denote by $E(c,i)$ this elementary matrix which multiplies the i^{th} row of the identity by the nonzero constant, c. Then from what was just discussed and the way matrices are multiplied,

$$E(c,i) \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ \vdots & \vdots & & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ip} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{np} \end{pmatrix}$$

equals a matrix having the columns indicated below.

$$= \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ \vdots & \vdots & & \vdots \\ ca_{i1} & ca_{i2} & \cdots & ca_{ip} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{np} \end{pmatrix}.$$

This proves the following lemma.

Lemma 1.6.4 *Let $E(c,i)$ denote the elementary matrix corresponding to the row operation in which the i^{th} row is multiplied by the nonzero constant c. Thus $E(c,i)$ involves multiplying the i^{th} row of the identity matrix by c. Then*

$$E(c,i)A = B$$

where B is obtained from A by multiplying the i^{th} row of A by c.

Finally consider the third of these row operations. Letting $\boldsymbol{r}_j$ be the j^{th} row of the identity matrix, denote by $E(c\times i+j)$ the elementary matrix obtained from the identity matrix by replacing $\boldsymbol{r}_j$ with $\boldsymbol{r}_j + c\boldsymbol{r}_i$. In case $i<j$ this will be of the form

$$\begin{pmatrix} \boldsymbol{r}_1 \\ \vdots \\ \boldsymbol{r}_i \\ \vdots \\ c\boldsymbol{r}_i + \boldsymbol{r}_j \\ \vdots \\ \boldsymbol{r}_n \end{pmatrix}.$$

Now consider what this does to a column vector.

$$\begin{pmatrix} \boldsymbol{r}_1 \\ \vdots \\ \boldsymbol{r}_i \\ \vdots \\ c\boldsymbol{r}_i + \boldsymbol{r}_j \\ \vdots \\ \boldsymbol{r}_n \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_i \\ \vdots \\ v_j \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} v_1 \\ \vdots \\ v_i \\ \vdots \\ cv_i + v_j \\ \vdots \\ v_n \end{pmatrix}.$$

Now from this and the way matrices are multiplied,

$$E(c\times i+j) \begin{pmatrix} a_{11} & a_{12} & \cdots & \cdots & \cdots & \cdots & a_{1p} \\ \vdots & \vdots & & & & & \vdots \\ a_{i1} & a_{i2} & \cdots & \cdots & \cdots & \cdots & a_{ip} \\ \vdots & \vdots & & & & & \vdots \\ a_{j2} & a_{j2} & \cdots & \cdots & \cdots & \cdots & a_{jp} \\ \vdots & \vdots & & & & & \vdots \\ a_{n1} & a_{n2} & \cdots & \cdots & \cdots & \cdots & a_{np} \end{pmatrix}$$

equals a matrix of the following form having the indicated columns.

$$\left(E(c\times i+j)\begin{pmatrix} a_{11}\\ \vdots\\ a_{i1}\\ \vdots\\ a_{j2}\\ \vdots\\ a_{n1}\end{pmatrix}, E(c\times i+j)\begin{pmatrix} a_{12}\\ \vdots\\ a_{i2}\\ \vdots\\ a_{j2}\\ \vdots\\ a_{n2}\end{pmatrix}, \cdots E(c\times i+j)\begin{pmatrix} a_{1p}\\ \vdots\\ a_{ip}\\ \vdots\\ a_{jp}\\ \vdots\\ a_{np}\end{pmatrix}\right)$$

$$= \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p}\\ \vdots & \vdots & & \vdots\\ a_{i1} & a_{i2} & \cdots & a_{ip}\\ \vdots & \vdots & & \vdots\\ a_{j2}+ca_{i1} & a_{j2}+ca_{i2} & \cdots & a_{jp}+ca_{ip}\\ \vdots & \vdots & & \vdots\\ a_{n1} & a_{n2} & \cdots & a_{np}\end{pmatrix}.$$

The case where $i > j$ is similar. This proves the following lemma in which, as above, the i^{th} row of the identity is $\boldsymbol{r}_i$.

Lemma 1.6.5 *Let $E(c\times i+j)$ denote the elementary matrix obtained from I by replacing the j^{th} row of the identity $\boldsymbol{r}_j$ with $c\boldsymbol{r}_i+\boldsymbol{r}_j$. Letting the k^{th} row of A be $\boldsymbol{a}_k$,*

$$E(c\times i+j)A = B$$

where B has the same rows as A except the j^{th} row of B is $c\boldsymbol{a}_i+\boldsymbol{a}_j$.

The above lemmas are summarized in the following theorem.

Theorem 1.6.6 *To perform any of the three row operations on a matrix A it suffices to do the row operation on the identity matrix, obtaining an elementary matrix E, and then take the product, EA. In addition to this, the following identities hold for the elementary matrices described above.*

$$E(c\times i+j)E(-c\times i+j) = E(-c\times i+j)E(c\times i+j) = I. \tag{1.8}$$

$$E(c,i)E\left(c^{-1},i\right) = E\left(c^{-1},i\right)E(c,i) = I. \tag{1.9}$$

$$P^{ij}P^{ij} = I. \tag{1.10}$$

Proof: Consider (1.8). Starting with I and taking $-c$ times the i^{th} row added to the j^{th} yields $E(-c\times i+j)$ which differs from I only in the j^{th} row. Now multiplying on the left by $E(c\times i+j)$ takes c times the i^{th} row and adds to the j^{th} thus restoring the j^{th} row to its original state. Thus $E(c\times i+j)E(-c\times i+j) = I$. Similarly $E(-c\times i+j)E(c\times i+j) = I$. The reasoning is similar for (1.9) and (1.10). ■

Each of these elementary matrices has a significant geometric significance. The effect of doing $E\left(\frac{1}{2}\times 3+1\right)$ shears the box in one direction. Of course there would be corresponding shears in the other directions also. Note that this does not change the volume. You should think about the geometric effect of the other elementary matrices on a box.

Definition 1.6.7 *For an $n \times n$ matrix A, an $n \times n$ matrix B which has the property that $AB = BA = I$ is denoted by A^{-1}. Such a matrix is called an **inverse.** When A has an inverse, it is called **invertible.***

The following lemma says that if a matrix acts like an inverse, then it is **the** inverse. Also, the product of invertible matrices is invertible.

Lemma 1.6.8 *If B, C are both inverses of A, then $B = C$. That is, there exists at most one inverse of a matrix. If $A_1, \cdots, A_m$ are each invertible $m \times m$ matrices, then the product $A_1 A_2 \cdots A_m$ is also invertible and*

$$(A_1 A_2 \cdots A_m)^{-1} = A_m^{-1} A_{m-1}^{-1} \cdots A_1^{-1}.$$

Proof. From the definition and associative law of matrix multiplication,

$$B = BI = B(AC) = (BA)C = IC = C.$$

This proves the uniqueness of the inverse.

Next suppose A, B are invertible. Then

$$AB\left(B^{-1}A^{-1}\right) = A\left(BB^{-1}\right)A^{-1} = AIA^{-1} = AA^{-1} = I$$

and also

$$\left(B^{-1}A^{-1}\right)AB = B^{-1}\left(A^{-1}A\right)B = B^{-1}IB = B^{-1}B = I.$$

It follows from Definition 1.6.7 that AB has an inverse and it is $B^{-1}A^{-1}$. Thus the case of $m = 1, 2$ in the claim of the lemma is true. Suppose this claim is true for k. Then

$$A_1 A_2 \cdots A_k A_{k+1} = (A_1 A_2 \cdots A_k) A_{k+1}.$$

By induction, the two matrices $(A_1 A_2 \cdots A_k)$, A_{k+1} are both invertible and

$$(A_1 A_2 \cdots A_k)^{-1} = A_k^{-1} \cdots A_2^{-1} A_1^{-1}.$$

By the case of the product of two invertible matrices shown above,

$$\begin{aligned} ((A_1 A_2 \cdots A_k) A_{k+1})^{-1} &= A_{k+1}^{-1} (A_1 A_2 \cdots A_k)^{-1} \\ &= A_{k+1}^{-1} A_k^{-1} \cdots A_2^{-1} A_1^{-1}. \end{aligned}$$

This proves the lemma. ■

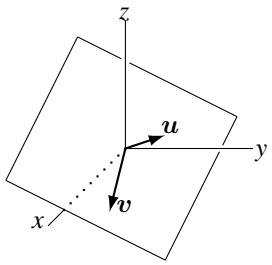
We will discuss methods for finding the inverse later. For now, observe that Theorem 1.6.6 says that elementary matrices are invertible and that the inverse of such a matrix is also an elementary matrix. The major conclusion of the above Lemma and Theorem is the following lemma about linear relationships.

Definition 1.6.9 *Let $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_k, \boldsymbol{u}$ be vectors. Then $\boldsymbol{u}$ is said to be a **linear combination** of the vectors $\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_k\}$ if there exist scalars $c_1, \cdots, c_k$ such that*

$$\boldsymbol{u} = \sum_{i=1}^{k} c_i \boldsymbol{v}_i.$$

*We also say that when the above holds for some scalars $c_1, \cdots, c_k$, there exists a **linear relationship** between the vector $\boldsymbol{u}$ and the vectors $\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_k\}$.*

We will discuss this more later, but the following picture illustrates the geometric significance of the vectors which have a linear relationship with two vectors $\boldsymbol{u},\boldsymbol{v}$ pointing in different directions.



The following lemma states that linear relationships between columns in a matrix are preserved by row operations. This simple lemma is the main result in understanding all the major questions related to the row reduced echelon form as well as many other topics.

Lemma 1.6.10 *Let A and B be two $m \times n$ matrices and suppose B results from a row operation applied to A. Then the k^{th} column of B is a linear combination of the $i_1, \cdots, i_r$ columns of B if and only if the k^{th} column of A is a linear combination of the $i_1, \cdots, i_r$ columns of A. Furthermore, the scalars in the linear combinations are the same. (The linear relationship between the k^{th} column of A and the $i_1, \cdots, i_r$ columns of A is the same as the linear relationship between the k^{th} column of B and the $i_1, \cdots, i_r$ columns of B.)*

Proof: Let A be the following matrix in which the $\boldsymbol{a}_k$ are the columns

$$\begin{pmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \cdots & \boldsymbol{a}_n \end{pmatrix}$$

and let B be the following matrix in which the columns are given by the $\boldsymbol{b}_k$

$$\begin{pmatrix} \boldsymbol{b}_1 & \boldsymbol{b}_2 & \cdots & \boldsymbol{b}_n \end{pmatrix}.$$

Then by Theorem 1.6.6 on Page 29, $\boldsymbol{b}_k = E\boldsymbol{a}_k$ where E is an elementary matrix. Suppose then that one of the columns of A is a linear combination of some other columns of A. Say

$$\boldsymbol{a}_k = c_1\boldsymbol{a}_{i_1} + \cdots + c_r\boldsymbol{a}_{i_r}.$$

Then multiplying by E,

$$\boldsymbol{b}_k = E\boldsymbol{a}_k = c_1E\boldsymbol{a}_{i_1} + \cdots + c_rE\boldsymbol{a}_{i_r} = c_1\boldsymbol{b}_{i_1} + \cdots + c_r\boldsymbol{b}_{i_r}.$$

This proves the lemma. ■

Example 1.6.11 Find linear relationships between the columns of the matrix

$$A = \begin{pmatrix} 1 & 3 & 11 & 10 & 36 \\ 1 & 2 & 8 & 9 & 23 \\ 1 & 1 & 5 & 8 & 10 \end{pmatrix}.$$

It is not clear what the relationships are, so we do row operations to this matrix. Lemma 1.6.10 says that all the linear relationships between columns are preserved, so the idea is to

do row operations until a matrix results which has the property that the linear relationships are obvious. First take -1 times the top row and add to the two bottom rows. This yields

$$\begin{pmatrix} 1 & 3 & 11 & 10 & 36 \\ 0 & -1 & -3 & -1 & -13 \\ 0 & -2 & -6 & -2 & -26 \end{pmatrix}$$

Next take -2 times the middle row and add to the bottom row followed by multiplying the middle row by -1 :

$$\begin{pmatrix} 1 & 3 & 11 & 10 & 36 \\ 0 & 1 & 3 & 1 & 13 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Next take -3 times the middle row added to the top:

$$\begin{pmatrix} 1 & 0 & 2 & 7 & -3 \\ 0 & 1 & 3 & 1 & 13 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{1.11}$$

At this point it is clear that the last column is -3 times the first column added to 13 times the second. By Lemma 1.6.10, the same is true of the corresponding columns in the original matrix A. As a check,

$$-3\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + 13\begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 36 \\ 23 \\ 10 \end{pmatrix}.$$

You should notice that other linear relationships are also easily seen from (1.11). For example the fourth column is 7 times the first added to the second. This is obvious from (1.11) and Lemma 1.6.10 says the same relationship holds for A.

This is really just an extension of the technique for finding solutions to a linear system of equations. In solving a system of equations earlier, row operations were used to exhibit the last column of an augmented matrix as a linear combination of the preceding columns. The **row reduced echelon form** just extends this by making obvious the linear relationships between **every** column, not just the last, and those columns preceding it. The matrix in (1.11) is in row reduced echelon form. The row reduced echelon form is the topic of the next section.

1.7 The Row Reduced Echelon Form Of A Matrix

When you do row operations on a matrix, there is an ultimate conclusion. It is called the **row reduced echelon form**. We show here that every matrix has such a row reduced echelon form and that this row reduced echelon form is unique. The significance is that it becomes possible to use the definite article in referring to **the** row reduced echelon form. Hence important conclusions about the original matrix may be logically deduced from an examination of its unique row reduced echelon form. First we need the following definition.

Definition 1.7.1 *Define special column vectors $\mathbf{e}_i$ as follows.*

$$\mathbf{e}_i = \begin{pmatrix} 0 & \cdots & 1 & \cdots & 0 \end{pmatrix}^T.$$

Recall that ^{T} says to take the transpose. Thus $\mathbf{e}_i$ is the column vector which has all zero entries except for a 1 in the i^{th} position down from the top.

Now here is the description of the row reduced echelon form.

Definition 1.7.2 *An $m \times n$ matrix is said to be in **row reduced echelon form** if, in viewing successive columns from left to right, the first nonzero column encountered is $\mathbf{e}_1$ and if, in viewing the columns of the matrix from left to right, you have encountered $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k$, the next column is either $\mathbf{e}_{k+1}$ or this next column is a linear combination of the vectors, $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k$.*

Example 1.7.3 The following matrices are in row reduced echelon form.

$$\begin{pmatrix} 1 & 0 & 4 & 0 \\ 0 & 1 & 3 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 & 0 & 7 \\ 0 & 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 & 3 \\ 0 & 0 & 1 & -5 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Definition 1.7.4 *Given a matrix A, row reduction produces one and only one row reduced matrix B with $A \sim B$. See Corollary 1.7.9. We call B **the** row reduced echelon form of A.*

Theorem 1.7.5 *Let A be an $m \times n$ matrix. Then A has a row reduced echelon form determined by a simple process.*

Proof. Viewing the columns of A from left to right, take the first nonzero column. Pick a nonzero entry in this column and switch the row containing this entry with the top row of A. Now divide this new top row by the value of this nonzero entry to get a 1 in this position and then use row operations to make all entries below this equal to zero. Thus the first nonzero column is now $\mathbf{e}_1$. Denote the resulting matrix by A_1. Consider the sub-matrix of A_1 to the right of this column and below the first row. Do exactly the same thing for this sub-matrix that was done for A. This time the $\mathbf{e}_1$ will refer to F^{m-1}. Use the first 1 obtained by the above process which is in the top row of this sub-matrix and row operations, to produce a zero in place of every entry above it and below it. Call the resulting matrix A_2. Thus A_2 satisfies the conditions of the above definition up to the column just encountered. Continue this way till every column has been dealt with and the result must be in row reduced echelon form. ■

Here is some terminology about pivot columns.

Definition 1.7.6 *The first **pivot column** of A is the first nonzero column of A which becomes $\mathbf{e}_1$ in the row reduced echelon form. The next pivot column is the first column after this which becomes $\mathbf{e}_2$ in the row reduced echelon form. The third is the next column which becomes $\mathbf{e}_3$ in the row reduced echelon form and so forth.*

The algorithm just described for obtaining a row reduced echelon form shows that these columns are well defined, but we will deal with this issue more carefully in Corollary 1.7.9 where we show that every matrix corresponds to exactly one row reduced echelon form.

Definition 1.7.7 *Two matrices A, B are said to be **row equivalent** if B can be obtained from A by a sequence of row operations. When A is row equivalent to B, we write $A \sim B$.*

Proposition 1.7.8 *In the notation of Definition 1.7.7. $A \sim A$. If $A \sim B$, then $B \sim A$. If $A \sim B$ and $B \sim C$, then $A \sim C$.*

Proof. That $A \sim A$ is obvious. Consider the second claim. By Theorem 1.6.6, there exist elementary matrices $E_1, E_2, \cdots, E_m$ such that

$$B = E_1 E_2 \cdots E_m A.$$

It follows from Lemma 1.6.8 that $(E_1 E_2 \cdots E_m)^{-1}$ exists and equals the product of the inverses of these matrices in the reverse order. Thus

$$E_m^{-1} E_{m-1}^{-1} \cdots E_1^{-1} B = (E_1 E_2 \cdots E_m)^{-1} B$$

$$= (E_1 E_2 \cdots E_m)^{-1} (E_1 E_2 \cdots E_m) A = A.$$

By Theorem 1.6.6, each E_k^{-1} is an elementary matrix. By Theorem 1.6.6 again, the above shows that A results from a sequence of row operations applied to B. The last claim is left for an exercise. This proves the proposition. ■

There are three choices for row operations at each step in Theorem 1.7.5. A natural question is whether the same row reduced echelon matrix always results in the end from following any sequence of row operations.

We have already made use of the following observation in finding a linear relationship between the columns of the matrix A, but here it is stated more formally.

$$\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = x_1 \mathbf{e}_1 + \cdots + x_n \mathbf{e}_n,$$

so to say two column vectors are equal, is to say the column vectors are the same linear combination of the special vectors $\mathbf{e}_j$.

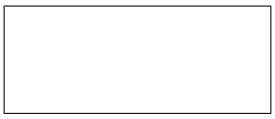
Corollary 1.7.9 *The row reduced echelon form is unique. That is if B, C are two matrices in row reduced echelon form and both are obtained from A by a sequence of row operations, then $B = C$.*

Proof. Suppose B and C are both row reduced echelon forms for the matrix A. It follows that B and C have zero columns in the same positions because row operations do not affect zero columns. By Proposition 1.7.8, B and C are row equivalent. In reading from left to right in B, suppose $\mathbf{e}_1, \cdots, \mathbf{e}_r$ occur first in positions $i_1, \cdots, i_r$ respectively. Then from the description of the row reduced echelon form, each of these columns of B, in positions $i_1, \cdots, i_r$, is not a linear combination of the preceding columns. Since C is row equivalent to B, it follows from Lemma 1.6.10, that each column of C in positions $i_1, \cdots, i_r$ is not a linear combination of the preceding columns of C. By the description of the row reduced echelon form, $\mathbf{e}_1, \cdots, \mathbf{e}_r$ occur first in C, in positions $i_1, \cdots, i_r$ respectively. Therefore, both B and C have the sequence $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_r$ occurring first (reading from left to right) in the positions, $i_1, i_2, \cdots, i_r$. Since these matrices are row equivalent, it follows from Lemma 1.6.10, that the columns between the i_k and i_{k+1} position in the two matrices are linear combinations involving the same scalars, of the columns in the $i_1, \cdots, i_k$ position. Similarly, the columns after the i_r position are linear combinations of the columns in the $i_1, \cdots, i_r$ positions involving the same scalars in both matrices. This is equivalent to the assertion that each of these columns is identical in B and C. ■

Now with the above corollary, here is a very fundamental observation. The number of nonzero rows in the row reduced echelon form is the same as the number of pivot columns.

Namely, this number is r in both cases where $\boldsymbol{e}_1, \cdots, \boldsymbol{e}_r$ are the pivot columns in the row reduced echelon form. This number r is called the **rank** of the matrix. This is discussed more later, but first here are some other applications.

Consider a matrix which looks like this: (More columns than rows.)



Corollary 1.7.10 *Suppose A is an $m \times n$ matrix and that $m < n$. That is, the number of rows is less than the number of columns. Then one of the columns of A is a linear combination of the preceding columns of A. Also, there exists $\boldsymbol{x} \in F^n$ such that $\boldsymbol{x} \neq \boldsymbol{0}$ and $A\boldsymbol{x} = \boldsymbol{0}$.*

Proof: Since $m < n$, not all the columns of A can be pivot columns. In reading from left to right, pick the first one which is not a pivot column. Then from the description of the row reduced echelon form, this column is a linear combination of the preceding columns. Say

$$\boldsymbol{a}_j = x_1\boldsymbol{a}_1 + \cdots + x_{j-1}\boldsymbol{a}_{j-1}.$$

Therefore, from the way we multiply a matrix times a vector,

$$A\begin{pmatrix} x_1 \\ \vdots \\ x_{j-1} \\ -1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \left(\boldsymbol{a}_1 \cdots \boldsymbol{a}_{j-1}\boldsymbol{a}_j \cdots \boldsymbol{a}_n\right)\begin{pmatrix} x_1 \\ \vdots \\ x_{j-1} \\ -1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \boldsymbol{0}. \ \blacksquare$$

1.8 Finding the Inverse of a Matrix

Recall that the inverse of an $n \times n$ matrix A is a matrix B such that

$$AB = BA = I$$

where I is the identity matrix. It was shown that an elementary matrix is invertible and that its inverse is also an elementary matrix. Also the product of invertible matrices is invertible and its inverse is the product of the inverses in the reverse order. In this section, we consider the problem of finding an inverse for a given $n \times n$ matrix.

Example 1.8.1 Let $A = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$. Show that $\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}$ is the inverse of A.

To check this, multiply

$$\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

and

$$\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

showing that this matrix is indeed the inverse of A.

In the last example, how would you find A^{-1}? You wish to find a matrix $\begin{pmatrix} x & z \\ y & w \end{pmatrix}$ such that

$$\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} x & z \\ y & w \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

This requires the solution of the systems of equations,

$$x+y=1, x+2y=0$$

and

$$z+w=0, z+2w=1.$$

Writing the augmented matrix for these two systems gives

$$\begin{pmatrix} 1 & 1 & | & 1 \\ 1 & 2 & | & 0 \end{pmatrix} \tag{1.12}$$

for the first system and

$$\begin{pmatrix} 1 & 1 & | & 0 \\ 1 & 2 & | & 1 \end{pmatrix} \tag{1.13}$$

for the second. Let's solve the first system. Take (-1) times the first row and add to the second to get

$$\begin{pmatrix} 1 & 1 & | & 1 \\ 0 & 1 & | & -1 \end{pmatrix}$$

Now take (-1) times the second row and add to the first to get

$$\begin{pmatrix} 1 & 0 & | & 2 \\ 0 & 1 & | & -1 \end{pmatrix}.$$

Putting in the variables, this says $x=2$ and $y=-1$.

Now solve the second system, (1.13) to find z and w. Take (-1) times the first row and add to the second to get

$$\begin{pmatrix} 1 & 1 & | & 0 \\ 0 & 1 & | & 1 \end{pmatrix}.$$

Now take (-1) times the second row and add to the first to get

$$\begin{pmatrix} 1 & 0 & | & -1 \\ 0 & 1 & | & 1 \end{pmatrix}.$$

Putting in the variables, this says $z=-1$ and $w=1$. Therefore, the inverse is

$$\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}.$$

Didn't the above seem rather repetitive? Exactly the same row operations were used in both systems. In each case, the end result was something of the form $(I|\boldsymbol{v})$ where I is the

identity and $\boldsymbol{v}$ gave a column of the inverse. In the above $\begin{pmatrix} x \\ y \end{pmatrix}$, the first column of the inverse was obtained first and then the second column $\begin{pmatrix} z \\ w \end{pmatrix}$.

To simplify this procedure, you could have written

$$\left(\begin{array}{cc|cc} 1 & 1 & 1 & 0 \\ 1 & 2 & 0 & 1 \end{array}\right)$$

and row reduced till you obtained

$$\left(\begin{array}{cc|cc} 1 & 0 & 2 & -1 \\ 0 & 1 & -1 & 1 \end{array}\right).$$

Then you could have read off the inverse as the 2×2 matrix on the right side. You should be able to see that it is valid by adapting the argument used in the simple case above.

This is the reason for the following simple procedure for finding the inverse of a matrix. This procedure is called the **Gauss-Jordan procedure**.

Procedure 1.8.2 *Suppose A is an $n\times n$ matrix. To find A^{-1} if it exists, form the augmented $n\times 2n$ matrix*

$$(A|I)$$

and then if possible, do row operations until you obtain an $n\times 2n$ matrix of the form

$$(I|B)\,. \tag{1.14}$$

When this has been done, $B=A^{-1}$. If it is impossible to row reduce to a matrix of the form $(I|B)$, then A has no inverse.

The procedure just described along with the preceding explanation shows that this procedure actually yields a **right inverse**. This is a matrix B such that $AB=I$. We will show in Theorem 1.8.4 that this right inverse is really **the** inverse. This is a stronger result than that of Lemma 1.6.8 about the uniqueness of the inverse. For now, here is an example.

Example 1.8.3 Let $A=\begin{pmatrix} 1 & 2 & 2 \\ 1 & 0 & 2 \\ 3 & 1 & -1 \end{pmatrix}$. Find A^{-1} if it exists.

Set up the augmented matrix $(A|I)$:

$$\left(\begin{array}{ccc|ccc} 1 & 2 & 2 & 1 & 0 & 0 \\ 1 & 0 & 2 & 0 & 1 & 0 \\ 3 & 1 & -1 & 0 & 0 & 1 \end{array}\right)$$

Next take (-1) times the first row and add to the second followed by (-3) times the first row added to the last. This yields

$$\left(\begin{array}{ccc|ccc} 1 & 2 & 2 & 1 & 0 & 0 \\ 0 & -2 & 0 & -1 & 1 & 0 \\ 0 & -5 & -7 & -3 & 0 & 1 \end{array}\right).$$

Then take 5 times the second row and add to -2 times the last row.

$$\left(\begin{array}{ccc|ccc} 1 & 2 & 2 & 1 & 0 & 0 \\ 0 & -10 & 0 & -5 & 5 & 0 \\ 0 & 0 & 14 & 1 & 5 & -2 \end{array}\right)$$

Next take the last row and add to (-7) times the top row. This yields

$$\left(\begin{array}{ccc|ccc} -7 & -14 & 0 & -6 & 5 & -2 \\ 0 & -10 & 0 & -5 & 5 & 0 \\ 0 & 0 & 14 & 1 & 5 & -2 \end{array}\right).$$

Now take $(-7/5)$ times the second row and add to the top.

$$\left(\begin{array}{ccc|ccc} -7 & 0 & 0 & 1 & -2 & -2 \\ 0 & -10 & 0 & -5 & 5 & 0 \\ 0 & 0 & 14 & 1 & 5 & -2 \end{array}\right).$$

Finally divide the top row by -7, the second row by -10 and the bottom row by 14, which yields

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 0 & -\frac{1}{7} & \frac{2}{7} & \frac{2}{7} \\ 0 & 1 & 0 & \frac{1}{2} & -\frac{1}{2} & 0 \\ 0 & 0 & 1 & \frac{1}{14} & \frac{5}{14} & -\frac{1}{7} \end{array}\right).$$

Therefore, the inverse is

$$\left(\begin{array}{ccc} -\frac{1}{7} & \frac{2}{7} & \frac{2}{7} \\ \frac{1}{2} & -\frac{1}{2} & 0 \\ \frac{1}{14} & \frac{5}{14} & -\frac{1}{7} \end{array}\right).$$

What you have really found in the above algorithm is a **right inverse.** Is this right inverse matrix, which we have called the inverse, really **the** inverse, the matrix which when multiplied on both sides gives the identity?

Theorem 1.8.4 *Suppose A, B are $n \times n$ matrices and $AB = I$. Then it follows that $BA = I$ also, and so $B = A^{-1}$. For $n \times n$ matrices, the left inverse, right inverse and inverse are all the same thing.*

Proof. If $AB = I$ for A, B $n \times n$ matrices, is $BA = I$? If $AB = I$, there exists a unique solution $\boldsymbol{x}$ to the equation

$$B\boldsymbol{x} = \boldsymbol{y}$$

for any choice of $\boldsymbol{y}$. In fact,

$$\boldsymbol{x} = A(B\boldsymbol{x}) = A\boldsymbol{y}.$$

This means the row reduced echelon form of B must be I. Thus every column is a pivot column. Otherwise, there exists a free variable and the solution, if it exists, would not be

unique, contrary to what was just shown must happen if $AB = I$. It follows that a right inverse B^{-1} for B exists. The above procedure yields

$$\left(\begin{array}{cc} B & I \end{array} \right) \rightarrow \left(\begin{array}{cc} I & B^{-1} \end{array} \right).$$

Now multiply both sides of the equation $AB = I$ on the right by B^{-1}. Then

$$A = A\left(BB^{-1}\right) = (AB)B^{-1} = B^{-1}.$$

Thus A is the right inverse of B, and so $BA = I$. This shows that if $AB = I$, then $BA = I$ also. Exchanging roles of A and B, we see that if $BA = I$, then $AB = I$. This proves the theorem. ■

This has shown that in the context of $n \times n$ matrices, right inverses, left inverses and inverses are all the same and this matrix is called A^{-1}.

The following corollary is also of interest.

Corollary 1.8.5 *An $n \times n$ matrix A has an inverse if and only if the row reduced echelon form of A is I.*

Proof. First suppose the row reduced echelon form of A is I. Then Procedure 1.8.2 yields a right inverse for A. By Theorem 1.8.4 this is **the** inverse. Next suppose A has an inverse. Then there exists a unique solution $\boldsymbol{x}$ to the equation $A\boldsymbol{x} = \boldsymbol{y}$. given by $\boldsymbol{x} = A^{-1}\boldsymbol{y}$. It follows that in the augmented matrix $(A|\mathbf{0})$ there are no free variables, and so every column to the left of the zero column is a pivot column. Therefore, the row reduced echelon form of A is I. ■

1.9 The Mathematical Theory of Determinants

It is easiest to give a different definition of the determinant which is clearly well defined and then prove the earlier one in terms of Laplace expansion. Let $(i_1, \cdots, i_n)$ be an ordered list of numbers from $\{1, \cdots, n\}$. This means the order is important so $(1,2,3)$ and $(2,1,3)$ are different. There will be some repetition between this section and the earlier section on determinants. The main purpose is to give all the missing proofs. Two books which give a good introduction to determinants are Apostol [1] and Rudin [48]. A recent book which also has a good introduction is Baker [4]

1.9.1 The Function sgn

The following Lemma will be essential in the definition of the determinant.

Lemma 1.9.1 *There exists a function, sgn_n which maps each ordered list of numbers from $\{1, \cdots, n\}$ to one of the three numbers, $0, 1$, or -1 which also has the following properties.*

$$\operatorname{sgn}_n(1, \cdots, n) = 1 \tag{1.15}$$

$$\operatorname{sgn}_n(i_1, \cdots, p, \cdots, q, \cdots, i_n) = -\operatorname{sgn}_n(i_1, \cdots, q, \cdots, p, \cdots, i_n) \tag{1.16}$$

In words, the second property states that if two of the numbers are switched, the value of the function is multiplied by -1. Also, in the case where $n > 1$ and $\{i_1, \cdots, i_n\} = \{1, \cdots, n\}$ so that every number from $\{1, \cdots, n\}$ appears in the ordered list, $(i_1, \cdots, i_n)$,

$$\operatorname{sgn}_n(i_1, \cdots, i_{\theta-1}, n, i_{\theta+1}, \cdots, i_n) \equiv$$

$$(-1)^{n-\theta}\,\mathrm{sgn}_{n-1}\left(i_1,\cdots,i_{\theta-1},i_{\theta+1},\cdots,i_n\right) \tag{1.17}$$

where $n = i_\theta$ in the ordered list, $(i_1,\cdots,i_n)$.

Proof: Define $\mathrm{sign}(x) = 1$ if $x > 0, -1$ if $x < 0$ and 0 if $x = 0$. If $n = 1$, there is only one list and it is just the number 1. Thus one can define $\mathrm{sgn}_1(1) \equiv 1$. For the general case where $n > 1$, simply define

$$\mathrm{sgn}_n\left(i_1,\cdots,i_n\right) \equiv \mathrm{sign}\left(\prod_{r<s}\left(i_s - i_r\right)\right)$$

This delivers either $-1, 1$, or 0 by definition. What about the other claims? Suppose you switch i_p with i_q where $p < q$ so two numbers in the ordered list $(i_1,\cdots,i_n)$ are switched. Denote the new ordered list of numbers as $(j_1,\cdots,j_n)$. Thus $j_p = i_q$ and $j_q = i_p$ and if $r \notin \{p,q\}$, $j_r = i_r$. See the following illustration

i_1	i_2	$\cdots$	i_p	$\cdots$	i_q	$\cdots$	i_n
1	2		p		q		n
i_1	i_2	$\cdots$	i_q	$\cdots$	i_p	$\cdots$	i_n
1	2		p		q		n
j_1	j_2	$\cdots$	j_p	$\cdots$	j_q	$\cdots$	j_n
1	2		p		q		n

Then

$$\mathrm{sgn}_n\left(j_1,\cdots,j_n\right) \equiv \mathrm{sign}\left(\prod_{r<s}\left(j_s - j_r\right)\right)$$

$$= \mathrm{sign}\left(\overset{\text{both } p,q}{\left(i_p - i_q\right)} \prod_{p<j<q}\left(i_j - i_q\right) \overset{\text{one of } p,q}{\prod_{p<j<q}}\left(i_p - i_j\right) \overset{\text{neither } p \text{ nor } q}{\prod_{r<s,r,s\notin\{p,q\}}}\left(i_s - i_r\right)\right)$$

The last product consists of the product of terms which were in $\prod_{r<s}(i_s - i_r)$ while the two products in the middle both introduce $q - p - 1$ minus signs. Thus their product is positive. The first factor is of opposite sign to the $i_q - i_p$ which occured in $\mathrm{sgn}_n(i_1,\cdots,i_n)$. Therefore, this switch introduced a minus sign and

$$\mathrm{sgn}_n\left(j_1,\cdots,j_n\right) = -\,\mathrm{sgn}_n\left(i_1,\cdots,i_n\right)$$

Now consider the last claim. In computing $\mathrm{sgn}_n(i_1,\cdots,i_{\theta-1},n,i_{\theta+1},\cdots,i_n)$ there will be the product of $n-\theta$ negative terms

$$\left(i_{\theta+1} - n\right)\cdots\left(i_n - n\right)$$

and the other terms in the product for computing $\mathrm{sgn}_n(i_1,\cdots,i_{\theta-1},n,i_{\theta+1},\cdots,i_n)$ are those which are required to compute $\mathrm{sgn}_{n-1}(i_1,\cdots,i_{\theta-1},i_{\theta+1},\cdots,i_n)$ multiplied by terms of the form $(n - i_j)$ which are nonnegative. It follows that

$$\mathrm{sgn}_n\left(i_1,\cdots,i_{\theta-1},n,i_{\theta+1},\cdots,i_n\right) = (-1)^{n-\theta}\,\mathrm{sgn}_{n-1}\left(i_1,\cdots,i_{\theta-1},i_{\theta+1},\cdots,i_n\right)$$

It is obvious that if there are repeats in the list the function gives 0. ■

Lemma 1.9.2 *Every ordered list of distinct numbers from* $\{1,2,\cdots,n\}$ *can be obtained from every other ordered list of distinct numbers by a finite number of switches. Also,* sgn_n *is unique.*

Proof: This is obvious if $n = 1$ or 2. Suppose then that it is true for sets of $n-1$ elements. Take two ordered lists of numbers, P_1, P_2. Make one switch in both to place n at the end. Call the result P_1^n and P_2^n. Then using induction, there are finitely many switches in P_1^n so that it will coincide with P_2^n. Now switch the n in what results to where it was in P_2.

To see sgn_n is unique, if there exist two functions, f and g both satisfying 1.15 and 1.16, you could start with $f(1,\cdots,n) = g(1,\cdots,n) = 1$ and applying the same sequence of switches, eventually arrive at $f(i_1,\cdots,i_n) = g(i_1,\cdots,i_n)$. If any numbers are repeated, then 1.16 gives both functions are equal to zero for that ordered list. ■

Definition 1.9.3 *Given an ordered list of distinct numbers from* $\{1,2,\cdots,n\}$, *say*

$$(i_1,\cdots,i_n),$$

this ordered list is called a permutation. The symbol for all such permutations is S_n. *The number* $\operatorname{sgn}_n(i_1,\cdots,i_n)$ *is called the sign of the permutation.*

A permutation can also be considered as a function from the set

$$\{1,2,\cdots,n\} \text{ to } \{1,2,\cdots,n\}$$

as follows. Let $f(k) = i_k$. Permutations are of fundamental importance in certain areas of math. For example, it was by considering permutations that Galois was able to give a criterion for solution of polynomial equations by radicals, but this is a different direction than what is being attempted here.

In what follows sgn will often be used rather than sgn_n because the context supplies the appropriate n.

1.9.2 The Definition of the Determinant

Definition 1.9.4 *Let* f *be a real valued function which has the set of ordered lists of numbers from* $\{1,\cdots,n\}$ *as its domain. Define*

$$\sum_{(k_1,\cdots,k_n)} f(k_1\cdots k_n)$$

to be the sum of all the $f(k_1\cdots k_n)$ *for all possible choices of ordered lists* $(k_1,\cdots,k_n)$ *of numbers of* $\{1,\cdots,n\}$. *For example,*

$$\sum_{(k_1,k_2)} f(k_1,k_2) = f(1,2) + f(2,1) + f(1,1) + f(2,2).$$

Definition 1.9.5 *Let* $(a_{ij}) = A$ *denote an* $n\times n$ *matrix. The determinant of* A, *denoted by* $\det(A)$ *is defined by*

$$\det(A) \equiv \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n)\, a_{1k_1}\cdots a_{nk_n}$$

where the sum is taken over all ordered lists of numbers from $\{1,\cdots,n\}$. *Note it suffices to take the sum over only those ordered lists in which there are no repeats because if there are,* $\operatorname{sgn}(k_1,\cdots,k_n)=0$ *and so that term contributes 0 to the sum.*

Let A be an $n\times n$ matrix $A=(a_{ij})$ and let $(r_1,\cdots,r_n)$ denote an ordered list of n numbers from $\{1,\cdots,n\}$. Let $A(r_1,\cdots,r_n)$ denote the matrix whose k^{th} row is the r_k row of the matrix A. Thus

$$\det(A(r_1,\cdots,r_n))=\sum_{(k_1,\cdots,k_n)}\operatorname{sgn}(k_1,\cdots,k_n)\,a_{r_1k_1}\cdots a_{r_nk_n} \tag{1.18}$$

and $A(1,\cdots,n)=A$.

Proposition 1.9.6 *Let* $(r_1,\cdots,r_n)$ *be an ordered list of numbers from*

$$\{1,\cdots,n\}$$

Then

$$\begin{aligned}\operatorname{sgn}(r_1,\cdots,r_n)\det(A) &= \sum_{(k_1,\cdots,k_n)}\operatorname{sgn}(k_1,\cdots,k_n)\,a_{r_1k_1}\cdots a_{r_nk_n} &(1.19)\\ &= \det(A(r_1,\cdots,r_n)). &(1.20)\end{aligned}$$

Proof: Let $(1,\cdots,n)=(1,\cdots,r,\cdots s,\cdots,n)$ so $r<s$.

$$\det(A(1,\cdots,r,\cdots,s,\cdots,n))= \tag{1.21}$$

$$\sum_{(k_1,\cdots,k_n)}\operatorname{sgn}(k_1,\cdots,k_r,\cdots,k_s,\cdots,k_n)\,a_{1k_1}\cdots a_{rk_r}\cdots a_{sk_s}\cdots a_{nk_n},$$

and renaming the variables, calling k_s,k_r and k_r, k_s, this equals

$$=\sum_{(k_1,\cdots,k_n)}\operatorname{sgn}(k_1,\cdots,k_s,\cdots,k_r,\cdots,k_n)\,a_{1k_1}\cdots a_{rk_s}\cdots a_{sk_r}\cdots a_{nk_n}$$

$$=\sum_{(k_1,\cdots,k_n)}-\operatorname{sgn}\left(k_1,\cdots,\overbrace{k_r,\cdots,k_s}^{\text{These got switched}},\cdots,k_n\right)a_{1k_1}\cdots a_{sk_r}\cdots a_{rk_s}\cdots a_{nk_n}$$

$$=-\det(A(1,\cdots,s,\cdots,r,\cdots,n)). \tag{1.22}$$

Consequently,

$$\det(A(1,\cdots,s,\cdots,r,\cdots,n))=-\det(A(1,\cdots,r,\cdots,s,\cdots,n))=-\det(A)$$

Now letting $A(1,\cdots,s,\cdots,r,\cdots,n)$ play the role of A, and continuing in this way, switching pairs of numbers,

$$\det(A(r_1,\cdots,r_n))=(-1)^p\det(A)$$

where it took p switches to obtain$(r_1,\cdots,r_n)$ from $(1,\cdots,n)$. By Lemma 1.9.1, this implies

$$\det(A(r_1,\cdots,r_n))=(-1)^p\det(A)=\operatorname{sgn}(r_1,\cdots,r_n)\det(A)$$

and proves the proposition in the case when there are no repeated numbers in the ordered list, $(r_1,\cdots,r_n)$. However, if there is a repeat, say the r^{th} row equals the s^{th} row, then the reasoning of 1.21 -1.22 shows that $\det(A(r_1,\cdots,r_n))=0$ and also $\operatorname{sgn}(r_1,\cdots,r_n)=0$ so the formula holds in this case also. ■

Observation 1.9.7 *There are $n!$ ordered lists of distinct numbers from*

$$\{1,\cdots,n\}$$

To see this, consider n slots placed in order. There are n choices for the first slot. For each of these choices, there are $n-1$ choices for the second. Thus there are $n(n-1)$ ways to fill the first two slots. Then for each of these ways there are $n-2$ choices left for the third slot. Continuing this way, there are $n!$ ordered lists of distinct numbers from $\{1,\cdots,n\}$ as stated in the observation.

1.9.3 A Symmetric Definition

With the above, it is possible to give a more symmetric description of the determinant from which it will follow that $\det(A)=\det\left(A^T\right)$.

Corollary 1.9.8 *The following formula for* $\det(A)$ *is valid.*

$$\det(A)=\frac{1}{n!}\cdot\sum_{(r_1,\cdots,r_n)}\sum_{(k_1,\cdots,k_n)}\operatorname{sgn}(r_1,\cdots,r_n)\operatorname{sgn}(k_1,\cdots,k_n)a_{r_1k_1}\cdots a_{r_nk_n}. \tag{1.23}$$

And also $\det\left(A^T\right)=\det(A)$ *where A^T is the transpose of A. (Recall that for $A^T=\left(a_{ij}^T\right)$, $a_{ij}^T=a_{ji}$.)*

Proof: From Proposition 1.9.6, if the r_i are distinct,

$$\det(A)=\sum_{(k_1,\cdots,k_n)}\operatorname{sgn}(r_1,\cdots,r_n)\operatorname{sgn}(k_1,\cdots,k_n)a_{r_1k_1}\cdots a_{r_nk_n}.$$

Summing over all ordered lists, $(r_1,\cdots,r_n)$ where the r_i are distinct, (If the r_i are not distinct, $\operatorname{sgn}(r_1,\cdots,r_n)=0$ and so there is no contribution to the sum.)

$$n!\det(A)=\sum_{(r_1,\cdots,r_n)}\sum_{(k_1,\cdots,k_n)}\operatorname{sgn}(r_1,\cdots,r_n)\operatorname{sgn}(k_1,\cdots,k_n)a_{r_1k_1}\cdots a_{r_nk_n}.$$

This proves the corollary since the formula gives the same number for A as it does for A^T. ■

Corollary 1.9.9 *If two rows or two columns in an $n\times n$ matrix A, are switched, the determinant of the resulting matrix equals (-1) times the determinant of the original matrix. If A is an $n\times n$ matrix in which two rows are equal or two columns are equal then* $\det(A)=0$. *Suppose the i^{th} row of A equals*

$$(xa_1+yb_1,\cdots,xa_n+yb_n)$$

Then

$$\det(A)=x\det(A_1)+y\det(A_2)$$

where the i^{th} row of A_1 is $(a_1,\cdots,a_n)$ and the i^{th} row of A_2 is $(b_1,\cdots,b_n)$, all other rows of A_1 and A_2 coinciding with those of A. In other words, det *is a linear function of each row A. The same is true with the word "row" replaced with the word "column".*

Proof: By Proposition 1.9.6 when two rows are switched, the determinant of the resulting matrix is (-1) times the determinant of the original matrix. By Corollary 1.9.8 the same holds for columns because the columns of the matrix equal the rows of the transposed matrix. Thus if A_1 is the matrix obtained from A by switching two columns,

$$\det(A) = \det\left(A^T\right) = -\det\left(A_1^T\right) = -\det(A_1).$$

If A has two equal columns or two equal rows, then switching them results in the same matrix. Therefore, $\det(A) = -\det(A)$ and so $\det(A) = 0$.

It remains to verify the last assertion.

$$\begin{aligned}\det(A) &\equiv \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n)\, a_{1k_1}\cdots\left(xa_{rk_i}+yb_{rk_i}\right)\cdots a_{nk_n}\\ &= x\sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n)\, a_{1k_1}\cdots a_{rk_i}\cdots a_{nk_n}\\ +y\sum_{(k_1,\cdots,k_n)} &\operatorname{sgn}(k_1,\cdots,k_n)\, a_{1k_1}\cdots b_{rk_i}\cdots a_{nk_n} \equiv x\det(A_1)+y\det(A_2).\end{aligned}$$

The same is true of columns because $\det\left(A^T\right) = \det(A)$ and the rows of A^T are the columns of A. ■

1.9.4 Basic Properties of the Determinant

Definition 1.9.10 *A vector, $\boldsymbol{w}$, is a linear combination $\{\boldsymbol{v}_1,\cdots,\boldsymbol{v}_r\}$ if there exist scalars $c_1,\cdots c_r$ such that $\boldsymbol{w} = \sum_{k=1}^r c_k\boldsymbol{v}_k$. This is the same as saying*

$$\boldsymbol{w} \in \operatorname{span}(\boldsymbol{v}_1,\cdots,\boldsymbol{v}_r).$$

The following corollary is also of great use.

Corollary 1.9.11 *Suppose A is an $n\times n$ matrix and some column (row) is a linear combination of r other columns (rows). Then* $\det(A) = 0$.

Proof: Let $A = \begin{pmatrix} \boldsymbol{a}_1 & \cdots & \boldsymbol{a}_n \end{pmatrix}$ be the columns of A and suppose the condition that one column is a linear combination of r of the others is satisfied. Say $\boldsymbol{a}_i = \sum_{j\neq i} c_j\boldsymbol{a}_j$. Then by Corollary 1.9.9, $\det(A) =$

$$\det\begin{pmatrix} \boldsymbol{a}_1 & \cdots & \sum_{j\neq i} c_j\boldsymbol{a}_j & \cdots & \boldsymbol{a}_n \end{pmatrix} = \sum_{j\neq i} c_j \det\begin{pmatrix} \boldsymbol{a}_1 & \cdots & \boldsymbol{a}_j & \cdots & \boldsymbol{a}_n \end{pmatrix} = 0$$

because each of these determinants in the sum has two equal rows. ■

Recall the following definition of matrix multiplication.

Definition 1.9.12 *If A and B are $n\times n$ matrices, $A = (a_{ij})$ and $B = (b_{ij})$, $AB = (c_{ij})$ where $c_{ij} \equiv \sum_{k=1}^n a_{ik}b_{kj}$.*

One of the most important rules about determinants is that the determinant of a product equals the product of the determinants.

Theorem 1.9.13 *Let A and B be $n \times n$ matrices. Then*

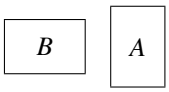
$$\det(AB) = \det(A)\det(B).$$

Proof: Let c_{ij} be the ij^{th} entry of AB. Then by Proposition 1.9.6,

$$\begin{aligned}\det(AB) &= \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n)\, c_{1k_1}\cdots c_{nk_n} \\ &= \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n)\left(\sum_{r_1} a_{1r_1}b_{r_1k_1}\right)\cdots\left(\sum_{r_n} a_{nr_n}b_{r_nk_n}\right)\end{aligned}$$

$$\begin{aligned}&= \sum_{(r_1\cdots,r_n)}\sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n)\, b_{r_1k_1}\cdots b_{r_nk_n}\left(a_{1r_1}\cdots a_{nr_n}\right) \\ &= \sum_{(r_1\cdots,r_n)} \operatorname{sgn}(r_1\cdots r_n)\, a_{1r_1}\cdots a_{nr_n}\det(B) = \det(A)\det(B).\blacksquare\end{aligned}$$

The Binet Cauchy formula is a generalization of the theorem which says the determinant of a product is the product of the determinants. The situation is illustrated in the following picture where A, B are matrices.



Theorem 1.9.14 *Let A be an $n \times m$ matrix with $n \geq m$ and let B be a $m \times n$ matrix. Also let A_i*

$$i = 1,\cdots,C(n,m)$$

be the $m \times m$ submatrices of A which are obtained by deleting $n-m$ rows and let B_i be the $m \times m$ submatrices of B which are obtained by deleting corresponding $n-m$ columns. Then

$$\det(BA) = \sum_{k=1}^{C(n,m)} \det(B_k)\det(A_k)$$

Proof: This follows from a computation. By Corollary 1.9.8 on Page 43, $\det(BA) =$

$$\frac{1}{m!}\sum_{(i_1\cdots i_m)}\sum_{(j_1\cdots j_m)} \operatorname{sgn}(i_1\cdots i_m)\operatorname{sgn}(j_1\cdots j_m)(BA)_{i_1j_1}(BA)_{i_2j_2}\cdots(BA)_{i_mj_m}$$

$$\begin{aligned}&= \frac{1}{m!}\sum_{(i_1\cdots i_m)}\sum_{(j_1\cdots j_m)} \operatorname{sgn}(i_1\cdots i_m)\operatorname{sgn}(j_1\cdots j_m)\cdot \\ &\sum_{r_1=1}^{n} B_{i_1r_1}A_{r_1j_1}\sum_{r_2=1}^{n} B_{i_2r_2}A_{r_2j_2}\cdots\sum_{r_m=1}^{n} B_{i_mr_m}A_{r_mj_m}\end{aligned}$$

Now denote by I_k one of the subsets of $\{1,\cdots,n\}$ which has m elements. Thus there are $C(n,m)$ of these.

$$\begin{aligned}&= \sum_{k=1}^{C(n,m)}\sum_{\{r_1,\cdots,r_m\}=I_k}\frac{1}{m!}\sum_{(i_1\cdots i_m)}\sum_{(j_1\cdots j_m)} \operatorname{sgn}(i_1\cdots i_m)\operatorname{sgn}(j_1\cdots j_m)\cdot \\ &B_{i_1r_1}A_{r_1j_1}B_{i_2r_2}A_{r_2j_2}\cdots B_{i_mr_m}A_{r_mj_m}\end{aligned}$$

$$= \sum_{k=1}^{C(n,m)} \sum_{\{r_1,\cdots,r_m\}=I_k} \frac{1}{m!} \sum_{(i_1\cdots i_m)} \operatorname{sgn}(i_1\cdots i_m) B_{i_1 r_1} B_{i_2 r_2} \cdots B_{i_m r_m} \cdot \sum_{(j_1\cdots j_m)} \operatorname{sgn}(j_1\cdots j_m) A_{r_1 j_1} A_{r_2 j_2} \cdots A_{r_m j_m}$$

$$= \sum_{k=1}^{C(n,m)} \sum_{\{r_1,\cdots,r_m\}=I_k} \frac{1}{m!} \operatorname{sgn}(r_1\cdots r_m)^2 \det(B_k)\det(A_k) = \sum_{k=1}^{C(n,m)} \det(B_k)\det(A_k)$$

since there are $m!$ ways of arranging the indices $\{r_1,\cdots,r_m\}$. ■

1.9.5 Expansion Using Cofactors

Lemma 1.9.15 *Suppose a matrix is of the form*

$$M = \begin{pmatrix} A & * \\ \mathbf{0} & a \end{pmatrix} \text{ or } \begin{pmatrix} A & \mathbf{0} \\ * & a \end{pmatrix} \tag{1.24}$$

where a is a number and A is an $(n-1)\times(n-1)$ matrix and $*$ denotes either a column or a row having length $n-1$ and the $\mathbf{0}$ denotes either a column or a row of length $n-1$ consisting entirely of zeros. Then $\det(M) = a\det(A)$.

Proof: Denote M by (m_{ij}). Thus in the first case, $m_{nn} = a$ and $m_{ni} = 0$ if $i \neq n$ while in the second case, $m_{nn} = a$ and $m_{in} = 0$ if $i \neq n$. From the definition of the determinant,

$$\det(M) \equiv \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}_n(k_1,\cdots,k_n) m_{1k_1}\cdots m_{nk_n}$$

Letting θ denote the position of n in the ordered list, $(k_1,\cdots,k_n)$ then using the earlier conventions used to prove Lemma 1.9.1, $\det(M)$ equals

$$\sum_{(k_1,\cdots,k_n)} (-1)^{n-\theta} \operatorname{sgn}_{n-1}\left(k_1,\cdots,k_{\theta-1},\overset{\theta}{k_{\theta+1}},\cdots,\overset{n-1}{k_n}\right) m_{1k_1}\cdots m_{nk_n}$$

Now suppose the second case. Then if $k_n \neq n$, the term involving m_{nk_n} in the above expression equals zero. Therefore, the only terms which survive are those for which $\theta = n$ or in other words, those for which $k_n = n$. Therefore, the above expression reduces to

$$a \sum_{(k_1,\cdots,k_{n-1})} \operatorname{sgn}_{n-1}(k_1,\cdots k_{n-1}) m_{1k_1}\cdots m_{(n-1)k_{n-1}} = a\det(A).$$

To get the assertion in the first case, use Corollary 1.9.8 to write

$$\det(M) = \det(M^T) = \det\left(\begin{pmatrix} A^T & \mathbf{0} \\ * & a \end{pmatrix}\right) = a\det(A^T) = a\det(A).\blacksquare$$

In terms of the theory of determinants, arguably the most important idea is that of Laplace expansion along a row or a column. This will follow from the above definition of a determinant.

Definition 1.9.16 *Let $A = (a_{ij})$ be an $n \times n$ matrix. Then a new matrix called the cofactor matrix* $\operatorname{cof}(A)$ *is defined by* $\operatorname{cof}(A) = (c_{ij})$ *where to obtain c_{ij} delete the i^{th} row and the j^{th} column of A, take the determinant of the $(n-1) \times (n-1)$ matrix which results, (This is called the ij^{th} minor of A.) and then multiply this number by $(-1)^{i+j}$. To make the formulas easier to remember,* $\operatorname{cof}(A)_{ij}$ *will denote the ij^{th} entry of the cofactor matrix.*

The following is the main result. Earlier this was given as a definition and the outrageous totally unjustified assertion was made that the same number would be obtained by expanding the determinant along any row or column. The following theorem proves this assertion.

Theorem 1.9.17 *Let A be an $n \times n$ matrix where $n \geq 2$. Then*

$$\det(A) = \sum_{j=1}^{n} a_{ij} \operatorname{cof}(A)_{ij} = \sum_{i=1}^{n} a_{ij} \operatorname{cof}(A)_{ij}. \tag{1.25}$$

The first formula consists of expanding the determinant along the i^{th} row and the second expands the determinant along the j^{th} column.

Proof: Let $(a_{i1}, \cdots, a_{in})$ be the i^{th} row of A. Let B_j be the matrix obtained from A by leaving every row the same except the i^{th} row which in B_j equals $(0, \cdots, 0, a_{ij}, 0, \cdots, 0)$. Then by Corollary 1.9.9,

$$\det(A) = \sum_{j=1}^{n} \det(B_j)$$

For example if

$$A = \begin{pmatrix} a & b & c \\ d & e & f \\ h & i & j \end{pmatrix}$$

and $i = 2$, then

$$B_1 = \begin{pmatrix} a & b & c \\ d & 0 & 0 \\ h & i & j \end{pmatrix}, B_2 = \begin{pmatrix} a & b & c \\ 0 & e & 0 \\ h & i & j \end{pmatrix}, B_3 = \begin{pmatrix} a & b & c \\ 0 & 0 & f \\ h & i & j \end{pmatrix}$$

Denote by A^{ij} the $(n-1) \times (n-1)$ matrix obtained by deleting the i^{th} row and the j^{th} column of A. Thus $\operatorname{cof}(A)_{ij} \equiv (-1)^{i+j} \det\left(A^{ij}\right)$. At this point, recall that from Proposition 1.9.6, when two rows or two columns in a matrix M, are switched, this results in multiplying the determinant of the old matrix by -1 to get the determinant of the new matrix. Therefore, by Lemma 1.9.15,

$$\begin{aligned} \det(B_j) &= (-1)^{n-j} (-1)^{n-i} \det\left(\begin{pmatrix} A^{ij} & * \\ \mathbf{0} & a_{ij} \end{pmatrix}\right) \\ &= (-1)^{i+j} \det\left(\begin{pmatrix} A^{ij} & * \\ \mathbf{0} & a_{ij} \end{pmatrix}\right) = a_{ij} \operatorname{cof}(A)_{ij}. \end{aligned}$$

Therefore,

$$\det(A) = \sum_{j=1}^{n} a_{ij} \operatorname{cof}(A)_{ij}$$

which is the formula for expanding $\det(A)$ along the i^{th} row. Also,

$$\det(A) = \det\left(A^T\right) = \sum_{j=1}^{n} a_{ij}^T \operatorname{cof}\left(A^T\right)_{ij} = \sum_{j=1}^{n} a_{ji} \operatorname{cof}(A)_{ji}$$

which is the formula for expanding $\det(A)$ along the i^{th} column. ■

1.9.6 A Formula for the Inverse

Note that this gives an easy way to write a formula for the inverse of an $n \times n$ matrix. Recall the definition of the inverse of a matrix in Definition 1.6.7 on Page 30.

Theorem 1.9.18 *A^{-1} exists if and only if $\det(A) \neq 0$. If $\det(A) \neq 0$, then $A^{-1} = \left(a_{ij}^{-1}\right)$ where*

$$a_{ij}^{-1} = \det(A)^{-1} \operatorname{cof}(A)_{ji}$$

for $\operatorname{cof}(A)_{ij}$ the ij^{th} cofactor of A.

Proof: By Theorem 1.9.17 and letting $(a_{ir}) = A$, if $\det(A) \neq 0$,

$$\sum_{i=1}^{n} a_{ir} \operatorname{cof}(A)_{ir} \det(A)^{-1} = \det(A) \det(A)^{-1} = 1.$$

Now in the matrix A, replace the k^{th} column with the r^{th} column and then expand along the k^{th} column. This yields for $k \neq r$,

$$\sum_{i=1}^{n} a_{ir} \operatorname{cof}(A)_{ik} \det(A)^{-1} = 0$$

because there are two equal columns by Corollary 1.9.9. Summarizing,

$$\sum_{i=1}^{n} a_{ir} \operatorname{cof}(A)_{ik} \det(A)^{-1} = \delta_{rk}.$$

Using the other formula in Theorem 1.9.17, and similar reasoning,

$$\sum_{j=1}^{n} a_{rj} \operatorname{cof}(A)_{kj} \det(A)^{-1} = \delta_{rk}$$

This proves that if $\det(A) \neq 0$, then A^{-1} exists with $A^{-1} = \left(a_{ij}^{-1}\right)$, where

$$a_{ij}^{-1} = \operatorname{cof}(A)_{ji} \det(A)^{-1}.$$

Now suppose A^{-1} exists. Then by Theorem 1.9.13,

$$1 = \det(I) = \det\left(AA^{-1}\right) = \det(A) \det\left(A^{-1}\right)$$

so $\det(A) \neq 0$. ■

The next corollary points out that if an $n \times n$ matrix A has a right or a left inverse, then it has an inverse.

Corollary 1.9.19 *Let A be an $n \times n$ matrix and suppose there exists an $n \times n$ matrix B such that $BA = I$. Then A^{-1} exists and $A^{-1} = B$. Also, if there exists C an $n \times n$ matrix such that $AC = I$, then A^{-1} exists and $A^{-1} = C$.*

Proof: Since $BA = I$, Theorem 1.9.13 implies $\det B \det A = 1$ and so $\det A \neq 0$. Therefore from Theorem 1.9.18, A^{-1} exists. Therefore,

$$A^{-1} = (BA)A^{-1} = B\left(AA^{-1}\right) = BI = B.$$

The case where $CA = I$ is handled similarly. ■

The conclusion of this corollary is that left inverses, right inverses and inverses are all the same in the context of $n \times n$ matrices.

Theorem 1.9.18 says that to find the inverse, take the transpose of the cofactor matrix and divide by the determinant. The transpose of the cofactor matrix is called the adjugate or sometimes the classical adjoint of the matrix A. It is an abomination to call it the adjoint although you do sometimes see it referred to in this way. In words, A^{-1} is equal to one over the determinant of A times the adjugate matrix of A.

1.9.7 Cramer's Rule

In case you are solving a system of equations, $A\boldsymbol{x} = \boldsymbol{y}$ for $\boldsymbol{x}$, it follows that if A^{-1} exists,

$$\boldsymbol{x} = \left(A^{-1}A\right)\boldsymbol{x} = A^{-1}\left(A\boldsymbol{x}\right) = A^{-1}\boldsymbol{y}$$

thus solving the system. Now in the case that A^{-1} exists, there is a formula for A^{-1} given above. Using this formula,

$$x_i = \sum_{j=1}^{n} a_{ij}^{-1} y_j = \sum_{j=1}^{n} \frac{1}{\det(A)} \operatorname{cof}(A)_{ji} y_j.$$

By the formula for the expansion of a determinant along a column,

$$x_i = \frac{1}{\det(A)} \det \begin{pmatrix} * & \cdots & y_1 & \cdots & * \\ \vdots & & \vdots & & \vdots \\ * & \cdots & y_n & \cdots & * \end{pmatrix},$$

where here the i^{th} column of A is replaced with the column vector, $(y_1 \cdots, y_n)^T$, and the determinant of this modified matrix is taken and divided by $\det(A)$. This formula is known as Cramer's rule.

Definition 1.9.20 *A matrix M, is upper triangular if $M_{ij} = 0$ whenever $i > j$. Thus such a matrix equals zero below the main diagonal, the entries of the form M_{ii} as shown.*

$$\begin{pmatrix} * & * & \cdots & * \\ 0 & * & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \cdots & 0 & * \end{pmatrix}$$

A lower triangular matrix is defined similarly as a matrix for which all entries above the main diagonal are equal to zero.

With this definition, here is a simple corollary of Theorem 1.9.17.

Corollary 1.9.21 *Let M be an upper (lower) triangular matrix. Then $\det(M)$ is obtained by taking the product of the entries on the main diagonal.*

1.9.8 Rank of a Matrix

Definition 1.9.22 *A submatrix of a matrix A is the rectangular array of numbers obtained by deleting some rows and columns of A. Let A be an $m \times n$ matrix. The **determinant rank** of the matrix equals r where r is the largest number such that some $r \times r$ submatrix of A has a non zero determinant. The **row rank** is defined to be the dimension of the span of the rows. The **column rank** is defined to be the dimension of the span of the columns.*

Theorem 1.9.23 *If A, an $m \times n$ matrix has determinant rank r, then there exist r rows of the matrix such that every other row is a linear combination of these r rows.*

Proof: Suppose the determinant rank of $A = (a_{ij})$ equals r. Thus some $r \times r$ submatrix has non zero determinant and there is no larger square submatrix which has non zero determinant. Suppose such a submatrix is determined by the r columns whose indices are

$$j_1 < \cdots < j_r$$

and the r rows whose indices are

$$i_1 < \cdots < i_r$$

I want to show that every row is a linear combination of these rows. Consider the l^{th} row and let p be an index between 1 and n. Form the following $(r+1) \times (r+1)$ matrix

$$\begin{pmatrix} a_{i_1 j_1} & \cdots & a_{i_1 j_r} & a_{i_1 p} \\ \vdots & & \vdots & \vdots \\ a_{i_r j_1} & \cdots & a_{i_r j_r} & a_{i_r p} \\ a_{l j_1} & \cdots & a_{l j_r} & a_{l p} \end{pmatrix}$$

Of course you can assume $l \notin \{i_1, \cdots, i_r\}$ because there is nothing to prove if the l^{th} row is one of the chosen ones. The above matrix has determinant 0. This is because if $p \notin \{j_1, \cdots, j_r\}$ then the above would be a submatrix of A which is too large to have non zero determinant. On the other hand, if $p \in \{j_1, \cdots, j_r\}$ then the above matrix has two columns which are equal so its determinant is still 0.

Expand the determinant of the above matrix along the last column. Let C_k denote the cofactor associated with the entry $a_{i_k p}$. This is not dependent on the choice of p. Remember, you delete the column and the row the entry is in and take the determinant of what is left and multiply by -1 raised to an appropriate power. Let C denote the cofactor associated with a_{lp}. This is given to be nonzero, it being the determinant of the matrix $r \times r$ matrix in the upper left corner. Thus $0 = a_{lp} C + \sum_{k=1}^{r} C_k a_{i_k p}$ which implies $a_{lp} = \sum_{k=1}^{r} \frac{-C_k}{C} a_{i_k p} \equiv \sum_{k=1}^{r} m_k a_{i_k p}$ Since this is true for every p and since m_k does not depend on p, this has shown the l^{th} row is a linear combination of the $i_1, i_2, \cdots, i_r$ rows. ■

Corollary 1.9.24 *The determinant rank equals the row rank.*

Proof: From Theorem 1.9.23, every row is in the span of r rows where r is the determinant rank. Therefore, the row rank (dimension of the span of the rows) is no larger than the determinant rank. Could the row rank be smaller than the determinant rank? If so, it follows from Theorem 1.9.23 that there exist p rows for $p < r \equiv$ determinant rank, such that the span of these p rows equals the row space. But then you could consider the $r \times r$ sub matrix which determines the determinant rank and it would follow that each of these rows would be in the span of the restrictions of the p rows just mentioned. By Theorem 4.2.3, the exchange theorem, the rows of this sub matrix would not be linearly independent and so some row is a linear combination of the others. By Corollary 1.9.11 the determinant would be 0, a contradiction. ■

Corollary 1.9.25 *If A has determinant rank r, then there exist r columns of the matrix such that every other column is a linear combination of these r columns. Also the column rank equals the determinant rank.*

Proof: This follows from the above by considering A^T. The rows of A^T are the columns of A and the determinant rank of A^T and A are the same. Therefore, from Corollary 1.9.24, column rank of A = row rank of A^T = determinant rank of A^T = determinant rank of A. ■

The following theorem is of fundamental importance and ties together many of the ideas presented above.

Theorem 1.9.26 *Let A be an $n \times n$ matrix. Then the following are equivalent.*

1. $\det(A) = 0$.
2. A, A^T *are not one to one.*
3. A *is not onto.*

Proof: Suppose $\det(A) = 0$. Then the determinant rank of $A = r < n$. Therefore, there exist r columns such that every other column is a linear combination of these columns by Theorem 1.9.23. In particular, it follows that for some m, the m^{th} column is a linear combination of all the others. Thus letting $A = \begin{pmatrix} \boldsymbol{a}_1 & \cdots & \boldsymbol{a}_m & \cdots & \boldsymbol{a}_n \end{pmatrix}$ where the columns are denoted by $\boldsymbol{a}_i$, there exists scalars α_i such that $\boldsymbol{a}_m = \sum_{k \neq m} \alpha_k \boldsymbol{a}_k$. Now consider the column vector, $\boldsymbol{x} \equiv \begin{pmatrix} \alpha_1 & \cdots & -1 & \cdots & \alpha_n \end{pmatrix}^T$. Then $A\boldsymbol{x} = -\boldsymbol{a}_m + \sum_{k \neq m} \alpha_k \boldsymbol{a}_k = \boldsymbol{0}$. Since also $A\boldsymbol{0} = \boldsymbol{0}$, it follows A is not one to one. Similarly, A^T is not one to one by the same argument applied to A^T. This verifies that 1.) implies 2.).

Now suppose 2.). Then since A^T is not one to one, it follows there exists $\boldsymbol{x} \neq \boldsymbol{0}$ such that $A^T\boldsymbol{x} = \boldsymbol{0}$. Taking the transpose of both sides yields $\boldsymbol{x}^T A = \boldsymbol{0}^T$ where the $\boldsymbol{0}^T$ is a $1 \times n$ matrix or row vector. Now if $A\boldsymbol{y} = \boldsymbol{x}$, then $|\boldsymbol{x}|^2 = \boldsymbol{x}^T (A\boldsymbol{y}) = (\boldsymbol{x}^T A)\boldsymbol{y} = \boldsymbol{0}\boldsymbol{y} = 0$ contrary to $\boldsymbol{x} \neq \boldsymbol{0}$. Consequently there can be no $\boldsymbol{y}$ such that $A\boldsymbol{y} = \boldsymbol{x}$ and so A is not onto. This shows that 2.) implies 3.).

Finally, suppose 3.). If 1.) does not hold, then $\det(A) \neq 0$ but then from Theorem 1.9.18 A^{-1} exists and so for every $\boldsymbol{y} \in \mathbb{F}^n$ there exists a unique $\boldsymbol{x} \in \mathbb{F}^n$ such that $A\boldsymbol{x} = \boldsymbol{y}$. In fact $\boldsymbol{x} = A^{-1}\boldsymbol{y}$. Thus A would be onto contrary to 3.). This shows 3.) implies 1.). ■

Corollary 1.9.27 *Let A be an $n \times n$ matrix. Then the following are equivalent.*

1. $det(A) \neq 0$.

2. *A and A^T are one to one.*

3. *A is onto.*

Proof: This follows immediately from the above theorem.

1.9.9 An Identity of Cauchy

Theorem 1.9.28 *Both the left and the right sides in the following yield the same polynomial in the variables a_i, b_i for $i \leq n$.*

$$\prod_{i,j}(a_i+b_j)\begin{vmatrix} \frac{1}{a_1+b_1} & \cdots & \frac{1}{a_1+b_n} \\ \vdots & & \vdots \\ \frac{1}{a_n+b_1} & \cdots & \frac{1}{a_n+b_n} \end{vmatrix} = \prod_{j<i}(a_i-a_j)(b_i-b_j). \tag{1.26}$$

Proof: The theorem is true if $n=2$. This follows from some computations. Suppose it is true for $n-1$, $n \geq 3$.

$$\begin{vmatrix} \frac{1}{a_1+b_1} & \frac{1}{a_1+b_2} & \cdots & \frac{1}{a_1+b_n} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{1}{a_{n-1}+b_1} & \frac{1}{a_{n-1}+b_2} & & \frac{1}{a_{n-1}+b_n} \\ \frac{1}{a_n+b_1} & \frac{1}{a_n+b_2} & \cdots & \frac{1}{a_n+b_n} \end{vmatrix}$$

$$= \begin{vmatrix} \frac{a_n-a_1}{(a_1+b_1)(b_1+a_n)} & \frac{a_n-a_1}{(a_1+b_2)(b_2+a_n)} & \cdots & \frac{a_n-a_1}{(a_1+b_n)(a_n+b_n)} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{a_n-a_{n-1}}{(a_{n-1}+b_1)(a_n+b_1)} & \frac{a_n-a_{n-1}}{(b_2+a_n)(b_2+a_{n-1})} & & \frac{a_n-a_{n-1}}{(a_n+b_n)(b_n+a_{n-1})} \\ \frac{1}{a_n+b_1} & \frac{1}{a_n+b_2} & \cdots & \frac{1}{a_n+b_n} \end{vmatrix}$$

Continuing to use the multilinear properties of determinants, this equals

$$\begin{vmatrix} \frac{1}{(a_1+b_1)(b_1+a_n)} & \frac{1}{(a_1+b_2)(b_2+a_n)} & \cdots & \frac{1}{(a_1+b_n)(a_n+b_n)} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{1}{(a_{n-1}+b_1)(a_n+b_1)} & \frac{1}{(b_2+a_n)(b_2+a_{n-1})} & & \frac{1}{(a_n+b_n)(b_n+a_{n-1})} \\ \frac{1}{a_n+b_1} & \frac{1}{a_n+b_2} & \cdots & \frac{1}{a_n+b_n} \end{vmatrix} \prod_{k=1}^{n-1}(a_n-a_k)$$

and this equals

$$\begin{vmatrix} \frac{1}{(a_1+b_1)} & \frac{1}{(a_1+b_2)} & \cdots & \frac{1}{(a_1+b_n)} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{1}{(a_{n-1}+b_1)} & \frac{1}{(b_2+a_{n-1})} & & \frac{1}{(b_n+a_{n-1})} \\ 1 & 1 & \cdots & 1 \end{vmatrix} \frac{\prod_{k=1}^{n-1}(a_n-a_k)}{\prod_{k=1}^{n}(a_n+b_k)}$$

Now take -1 times the last column and add to each previous column. Thus it equals

$$\begin{vmatrix} \frac{b_n-b_1}{(a_1+b_1)(a_1+b_n)} & \frac{b_n-b_2}{(a_1+b_2)(a_1+b_n)} & \cdots & \frac{1}{(a_1+b_n)} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{b_n-b_1}{(b_1+a_{n-1})(b_n+a_{n-1})} & \frac{b_n-b_2}{(b_2+a_{n-1})(b_n+a_{n-1})} & & \frac{1}{(a_{n-1}+b_n)} \\ 0 & 0 & \cdots & 1 \end{vmatrix} \frac{\prod_{k=1}^{n-1}(a_n-a_k)}{\prod_{k=1}^{n}(a_n+b_k)}$$

Now continue simplifying using the multilinear property of the determinant.

$$\begin{vmatrix} \frac{1}{(a_1+b_1)} & \frac{1}{(a_1+b_2)} & \cdots & 1 \\ \vdots & \vdots & \cdots & \vdots \\ \frac{1}{(b_1+a_{n-1})} & \frac{1}{(b_2+a_{n-1})} & & 1 \\ 0 & 0 & \cdots & 1 \end{vmatrix} \frac{\prod_{k=1}^{n-1}(a_n-a_k)}{\prod_{k=1}^{n}(a_n+b_k)} \frac{\prod_{k=1}^{n-1}(b_n-b_k)}{\prod_{k=1}^{n-1}(a_k+b_n)}$$

Now, expanding along the bottom row, what has just resulted is

$$\begin{vmatrix} \frac{1}{a_1+b_1} & \cdots & \frac{1}{a_1+b_{n-1}} \\ \vdots & \cdots & \vdots \\ \frac{1}{a_{n-1}+b_1} & \cdots & \frac{1}{a_{n-1}+b_{n-1}} \end{vmatrix} \frac{\prod_{k=1}^{n-1}(a_n-a_k)}{\prod_{k=1}^{n}(a_n+b_k)} \frac{\prod_{k=1}^{n-1}(b_n-b_k)}{\prod_{k=1}^{n-1}(a_k+b_n)}$$

By induction this equals

$$\frac{\prod_{k=1}^{n-1}(a_n-a_k)}{\prod_{k=1}^{n}(a_n+b_k)} \frac{\prod_{k=1}^{n-1}(b_n-b_k)}{\prod_{k=1}^{n-1}(a_k+b_n)} \frac{\prod_{j<i\leq n-1}(a_i-a_j)(b_i-b_j)}{\prod_{i,j\leq n-1}(a_i+b_j)}$$

$$= \frac{\prod_{j<i\leq n}(a_i-a_j)(b_i-b_j)}{\prod_{i,j\leq n}(a_i+b_j)} \blacksquare$$

1.10 The Cayley Hamilton Theorem

Definition 1.10.1 *Let A be an $n\times n$ matrix. The characteristic polynomial is defined as*

$$q_A(t) \equiv \det(tI-A)$$

and the solutions to $q_A(t)=0$ are called eigenvalues. For A a matrix and $p(t)=t^n+a_{n-1}t^{n-1}+\cdots+a_1t+a_0$, denote by $p(A)$ the matrix defined by

$$p(A) \equiv A^n+a_{n-1}A^{n-1}+\cdots+a_1A+a_0I.$$

The explanation for the last term is that A^0 is interpreted as I, the identity matrix.

The Cayley Hamilton theorem states that every matrix satisfies its characteristic equation, that equation defined by $q_A(t)=0$. It is one of the most important theorems in linear algebra[2]. The proof in this section is not the most general proof, but works well when the field of scalars is $\mathbb{R}$ or $\mathbb{C}$. The following lemma will help with its proof.

Lemma 1.10.2 *Suppose for all $|\lambda|$ large enough,*

$$A_0+A_1\lambda+\cdots+A_m\lambda^m=0,$$

where the A_i are $n\times n$ matrices. Then each $A_i=0$.

[2]A special case was first proved by Hamilton in 1853. The general case was announced by Cayley some time later and a proof was given by Frobenius in 1878.

Proof: Suppose some $A_i \neq 0$. Let p be the largest index of those which are non zero. Then multiply by λ^{-p}.

$$A_0\lambda^{-p} + A_1\lambda^{-p+1} + \cdots + A_{p-1}\lambda^{-1} + A_p = 0$$

Now let $\lambda \to \infty$. Thus $A_p = 0$ after all. Hence each $A_i = 0$. ■

With the lemma, here is a simple corollary.

Corollary 1.10.3 *Let A_i and B_i be $n \times n$ matrices and suppose*

$$A_0 + A_1\lambda + \cdots + A_m\lambda^m = B_0 + B_1\lambda + \cdots + B_m\lambda^m$$

for all $|\lambda|$ large enough. Then $A_i = B_i$ for all i. If $A_i = B_i$ for each A_i, B_i then one can substitute an $n \times n$ matrix M for λ and the identity will continue to hold.

Proof: Subtract and use the result of the lemma. The last claim is obvious by matching terms. ■

With this preparation, here is a relatively easy proof of the Cayley Hamilton theorem.

Theorem 1.10.4 *Let A be an $n \times n$ matrix and let $q(\lambda) \equiv \det(\lambda I - A)$ be the characteristic polynomial. Then $q(A) = 0$.*

Proof: Let $C(\lambda)$ equal the transpose of the cofactor matrix of $(\lambda I - A)$ for $|\lambda|$ large. (If $|\lambda|$ is large enough, then λ cannot be in the finite list of eigenvalues of A and so for such λ, $(\lambda I - A)^{-1}$ exists.) Therefore, by Theorem 1.9.18

$$C(\lambda) = q(\lambda)(\lambda I - A)^{-1}.$$

Say

$$q(\lambda) = a_0 + a_1\lambda + \cdots + \lambda^n$$

Note that each entry in $C(\lambda)$ is a polynomial in λ having degree no more than $n-1$. For example, you might have something like

$$C(\lambda) = \begin{pmatrix} \lambda^2 - 6\lambda + 9 & 3 - \lambda & 0 \\ 2\lambda - 6 & \lambda^2 - 3\lambda & 0 \\ \lambda - 1 & \lambda - 1 & \lambda^2 - 3\lambda + 2 \end{pmatrix}$$

$$= \begin{pmatrix} 9 & 3 & 0 \\ -6 & 0 & 0 \\ -1 & -1 & 2 \end{pmatrix} + \lambda \begin{pmatrix} -6 & -1 & 0 \\ 2 & -3 & 0 \\ 1 & 1 & -3 \end{pmatrix} + \lambda^2 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Therefore, collecting the terms in the general case,

$$C(\lambda) = C_0 + C_1\lambda + \cdots + C_{n-1}\lambda^{n-1}$$

for C_j some $n \times n$ matrix. Then

$$C(\lambda)(\lambda I - A) = \left(C_0 + C_1\lambda + \cdots + C_{n-1}\lambda^{n-1}\right)(\lambda I - A) = q(\lambda) I$$

Then multiplying out the middle term, it follows that for all $|\lambda|$ sufficiently large,

$$a_0 I + a_1 I\lambda + \cdots + I\lambda^n = C_0\lambda + C_1\lambda^2 + \cdots + C_{n-1}\lambda^n$$

$$- \left[C_0 A + C_1 A\lambda + \cdots + C_{n-1}A\lambda^{n-1}\right]$$

$$= -C_0 A + (C_0 - C_1 A)\lambda + (C_1 - C_2 A)\lambda^2 + \cdots + (C_{n-2} - C_{n-1}A)\lambda^{n-1} + C_{n-1}\lambda^n$$

Then, using Corollary 1.10.3, one can replace λ on both sides with A. Then the right side is seen to equal 0. Hence the left side, $q(A)I$ is also equal to 0. ■

Part I

Topology, Continuity, Algebra, Derivatives

Chapter 2

Some Basic Topics

This chapter contains basic definitions and a few fundamental theorems which will be used throughout the book whenever convenient.

2.1 Basic Definitions

A set is a collection of things called elements of the set. For example, the set of integers, the collection of signed whole numbers such as 1,2,−4, etc. This set whose existence will be assumed is denoted by $\mathbb{Z}$. Other sets could be the set of people in a family or the set of donuts in a display case at the store. Sometimes parentheses, $\{\ \}$ specify a set by listing the things which are in the set between the parentheses. For example the set of integers between -1 and 2, including these numbers could be denoted as $\{-1,0,1,2\}$. The notation signifying x is an element of a set S, is written as $x \in S$. Thus, $1 \in \{-1,0,1,2,3\}$. Here are some axioms about sets. Axioms are statements which are accepted, not proved.

Axiom 2.1.1 *Two sets are equal if and only if they have the same elements.*

Axiom 2.1.2 *To every set, A, and to every condition $S(x)$ there corresponds a set, B, whose elements are exactly those elements x of A for which $S(x)$ holds.*

Axiom 2.1.3 *For every collection of sets there exists a set that contains all the elements that belong to at least one set of the given collection. (You can take the union of a bunch of sets.)*

Axiom 2.1.4 *The Cartesian product of a nonempty family of nonempty sets is nonempty.*

Axiom 2.1.5 *If A is a set there exists a set, $\mathscr{P}(A)$ such that $\mathscr{P}(A)$ is the set of all subsets of A. This is called the power set.*

These axioms are referred to as the axiom of extension, axiom of specification, axiom of unions, axiom of choice, and axiom of powers respectively.

It seems fairly clear you should want to believe in the axiom of extension. It is merely saying, for example, that $\{1,2,3\} = \{2,3,1\}$ since these two sets have the same elements in them. Similarly, it would seem you should be able to specify a new set from a given set using some "condition" which can be used as a test to determine whether the element in question is in the set. For example, the set of all integers which are multiples of 2. This set could be specified as follows.

$$\{x \in \mathbb{Z} : x = 2y \text{ for some } y \in \mathbb{Z}\}.$$

In this notation, the colon is read as "such that" and in this case the condition is being a multiple of 2.

Another example of political interest, could be the set of all judges who are not judicial activists. I think you can see this last is not a very precise condition since there is no way to determine to everyone's satisfaction whether a given judge is an activist. Also, **just because something is grammatically correct does not mean it makes any sense.** For example consider the following nonsense.

$$S = \{x \in \text{set of dogs} : \text{it is colder in the mountains than in the winter}\}.$$

So what is a condition?

We will leave these sorts of considerations and assume our conditions make sense, whatever that means. The axiom of unions states that for any collection of sets, there is a set consisting of all the elements in each of the sets in the collection. Of course this is also open to further consideration. What is a collection? Maybe it would be better to say "set of sets" or, given a set whose elements are sets there exists a set whose elements consist of exactly those things which are elements of at least one of these sets. If $\mathscr{S}$ is such a set whose elements are sets,

$$\cup\{A : A \in \mathscr{S}\} \text{ or } \cup\mathscr{S}$$

signify this union.

Something is in the Cartesian product of a set or "family" of sets if it consists of a single thing taken from each set in the family. Thus $(1,2,3) \in \{1,4,.2\} \times \{1,2,7\} \times \{4,3,7,9\}$ because it consists of exactly one element from each of the sets which are separated by $\times$. Also, this is the notation for the Cartesian product of finitely many sets. If $\mathscr{S}$ is a set whose elements are sets, $\prod_{A\in\mathscr{S}} A$ signifies the Cartesian product.

The Cartesian product is the set of choice functions, a choice function being a function which selects exactly one element of each set of $\mathscr{S}$. You may think the axiom of choice, stating that the Cartesian product of a nonempty family of nonempty sets is nonempty, is innocuous but there was a time when many mathematicians were ready to throw it out because it implies things which are very hard to believe, things which never happen without the axiom of choice.

A is a subset of B, written $A \subseteq B$, if every element of A is also an element of B. This can also be written as $B \supseteq A$. A is a proper subset of B, written $A \subset B$ or $B \supset A$ if A is a subset of B but A is not equal to B,$A \neq B$. $A \cap B$ denotes the intersection of the two sets, A and B and it means the set of elements of A which are also elements of B. The axiom of specification shows this is a set. The empty set is the set which has no elements in it, denoted as $\emptyset$. $A \cup B$ denotes the union of the two sets, A and B and it means the set of all elements which are in either of the sets. It is a set because of the axiom of unions.

The complement of a set, (the set of things which are not in the given set) must be taken with respect to a given set called the universal set which is a set which contains the one whose complement is being taken. Thus, the complement of A, denoted as A^C (or more precisely as $X \setminus A$) is a set obtained from using the axiom of specification to write

$$A^C \equiv \{x \in X : x \notin A\}$$

The symbol $\notin$ means: "is not an element of". Note the axiom of specification takes place relative to a given set. Without this universal set it makes no sense to use the axiom of specification to obtain the complement.

Words such as "all" or "there exists" are called quantifiers and they must be understood relative to some given set. For example, the set of all integers larger than 3. Or there exists an integer larger than 7. Such statements have to do with a given set, in this case the integers. Failure to have a reference set when quantifiers are used turns out to be illogical even though such usage may be grammatically correct. Quantifiers are used often enough that there are symbols for them. The symbol $\forall$ is read as "for all" or "for every" and the symbol $\exists$ is read as "there exists". Thus $\forall\forall\exists\exists$ could mean for every upside down A there exists a backwards E.

DeMorgan's laws are very useful in mathematics. Let $\mathscr{S}$ be a set of sets each of which

is contained in some universal set, U. Then

$$\cup\left\{A^C : A \in \mathscr{S}\right\} = \left(\cap\left\{A : A \in \mathscr{S}\right\}\right)^C$$

and

$$\cap\left\{A^C : A \in \mathscr{S}\right\} = \left(\cup\left\{A : A \in \mathscr{S}\right\}\right)^C.$$

These laws follow directly from the definitions. Also following directly from the definitions are:

Let $\mathscr{S}$ be a set of sets then

$$B \cup \cup\left\{A : A \in \mathscr{S}\right\} = \cup\left\{B \cup A : A \in \mathscr{S}\right\}.$$

and: Let $\mathscr{S}$ be a set of sets show

$$B \cap \cup\left\{A : A \in \mathscr{S}\right\} = \cup\left\{B \cap A : A \in \mathscr{S}\right\}.$$

Unfortunately, there is no single universal set which can be used for all sets. Here is why: Suppose there were. Call it S. Then you could consider A the set of all elements of S which are not elements of themselves, this from the axiom of specification. If A is an element of itself, then it fails to qualify for inclusion in A. Therefore, it must not be an element of itself. However, if this is so, it qualifies for inclusion in A so it is an element of itself and so this can't be true either. Thus the most basic of conditions you could imagine, that of being an element of, is meaningless and so allowing such a set causes the whole theory to be meaningless. The solution is to not allow a universal set. As mentioned by Halmos in Naive set theory, "Nothing contains everything". Always beware of statements involving quantifiers wherever they occur, even this one. This little observation described above is due to Bertrand Russell and is called Russell's paradox.

2.2 The Schroder Bernstein Theorem

It is very important to be able to compare the size of sets in a rational way. The most useful theorem in this context is the Schroder Bernstein theorem which is the main result to be presented in this section. The Cartesian product is discussed above. The next definition reviews this and defines the concept of a function.

Definition 2.2.1 *Let X and Y be sets.*

$$X \times Y \equiv \{(x,y) : x \in X \text{ and } y \in Y\}$$

A relation is defined to be a subset of $X \times Y$. A function f, also called a mapping, is a relation which has the property that if (x,y) and (x,y_1) are both elements of the f, then $y = y_1$. The domain of f is defined as

$$D(f) \equiv \{x : (x,y) \in f\},$$

written as $f : D(f) \to Y$. Another notation which is used is the following

$$f^{-1}(y) \equiv \{x \in D(f) : f(x) = y\}$$

This is called the inverse image.

It is probably safe to say that most people do not think of functions as a type of relation which is a subset of the Cartesian product of two sets. A function is like a machine which takes inputs, x and makes them into a unique output, $f(x)$. Of course, that is what the above definition says with more precision. An ordered pair, (x,y) which is an element of the function or mapping has an input, x and a unique output y,denoted as $f(x)$ while the name of the function is f. "mapping" is often a noun meaning function. However, it also is a verb as in "f is mapping A to B". That which a function is thought of as doing is also referred to using the word "maps" as in: f maps X to Y. However, a set of functions may be called a set of maps so this word might also be used as the plural of a noun. There is no help for it. You just have to suffer with this nonsense.

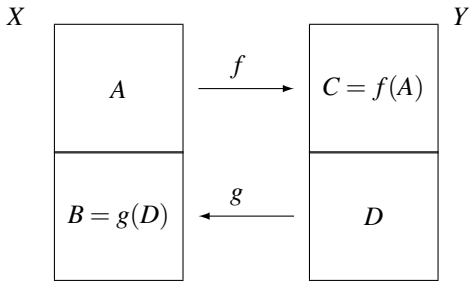
The following theorem which is interesting for its own sake will be used to prove the Schroder Bernstein theorem.

Theorem 2.2.2 *Let $f : X \to Y$ and $g : Y \to X$ be two functions. Then there exist sets A,B,C,D, such that*

$$A \cup B = X,\ C \cup D = Y,\ A \cap B = \emptyset,\ C \cap D = \emptyset,$$

$$f(A) = C,\ g(D) = B.$$

The following picture illustrates the conclusion of this theorem.



Proof: Consider the empty set, $\emptyset \subseteq X$. If $y \in Y \setminus f(\emptyset)$, then $g(y) \notin \emptyset$ because $\emptyset$ has no elements. Also, if A,B,C, and D are as described above, A also would have this same property that the empty set has. However, A is probably larger. Therefore, say $A_0 \subseteq X$ satisfies $\mathscr{P}$ if whenever $y \in Y \setminus f(A_0)$, $g(y) \notin A_0$.

$$\mathscr{A} \equiv \{A_0 \subseteq X : A_0 \text{ satisfies } \mathscr{P}\}.$$

Let $A = \cup\mathscr{A}$. If $y \in Y \setminus f(A)$, then for each $A_0 \in \mathscr{A}$, $y \in Y \setminus f(A_0)$ and so $g(y) \notin A_0$. Since $g(y) \notin A_0$ for all $A_0 \in \mathscr{A}$, it follows $g(y) \notin A$. Hence A satisfies $\mathscr{P}$ and is the largest subset of X which does so. Now define

$$C \equiv f(A),\ D \equiv Y \setminus C,\ B \equiv X \setminus A.$$

It only remains to verify that $g(D) = B$. It was just shown that $g(D) \subseteq B$.

Suppose $x \in B = X \setminus A$. Then $A \cup \{x\}$ does not satisfy $\mathscr{P}$ and so there exists $y \in Y \setminus f(A \cup \{x\}) \subseteq D$ such that $g(y) \in A \cup \{x\}$. But $y \notin f(A)$ and so since A satisfies $\mathscr{P}$, it follows $g(y) \notin A$. Hence $g(y) = x$ and so $x \in g(D)$. Hence $g(D) = B$. ■

Theorem 2.2.3 *(Schroder Bernstein) If $f : X \to Y$ and $g : Y \to X$ are one to one, then there exists $h : X \to Y$ which is one to one and onto.*

Proof: Let A,B,C,D be the sets of Theorem2.2.2 and define

$$h(x) \equiv \begin{cases} f(x) & \text{if } x \in A \\ g^{-1}(x) & \text{if } x \in B \end{cases}$$

Then h is the desired one to one and onto mapping. ■

Recall that the Cartesian product may be considered as the collection of choice functions.

Definition 2.2.4 *Let I be a set and let X_i be a set for each $i \in I$. f is a choice function written as*

$$f \in \prod_{i \in I} X_i$$

if $f(i) \in X_i$ for each $i \in I$.

The axiom of choice says that if $X_i \neq \emptyset$ for each $i \in I$, for I a set, then

$$\prod_{i \in I} X_i \neq \emptyset.$$

Sometimes the two functions, f and g are onto but not one to one. It turns out that with the axiom of choice, a similar conclusion to the above may be obtained.

Corollary 2.2.5 *If $f : X \to Y$ is onto and $g : Y \to X$ is onto, then there exists $h : X \to Y$ which is one to one and onto.*

Proof: For each $y \in Y$, $f^{-1}(y) \equiv \{x \in X : f(x) = y\} \neq \emptyset$. Therefore, by the axiom of choice, there exists $f_0^{-1} \in \prod_{y \in Y} f^{-1}(y)$ which is the same as saying that for each $y \in Y$, $f_0^{-1}(y) \in f^{-1}(y)$. Similarly, there exists $g_0^{-1}(x) \in g^{-1}(x)$ for all $x \in X$. Then f_0^{-1} is one to one because if $f_0^{-1}(y_1) = f_0^{-1}(y_2)$, then

$$y_1 = f\left(f_0^{-1}(y_1)\right) = f\left(f_0^{-1}(y_2)\right) = y_2.$$

Similarly g_0^{-1} is one to one. Therefore, by the Schroder Bernstein theorem, there exists $h : X \to Y$ which is one to one and onto. ■

Definition 2.2.6 *A set S, is finite if there exists a natural number n and a map θ which maps $\{1,\cdots,n\}$ one to one and onto S. S is infinite if it is not finite. A set S, is called countable if there exists a map θ mapping $\mathbb{N}$ one to one and onto S.(When θ maps a set A to a set B, this will be written as $\theta : A \to B$ in the future.) Here $\mathbb{N} \equiv \{1,2,\cdots\}$, the natural numbers. S is at most countable if there exists a map $\theta : \mathbb{N} \to S$ which is onto.*

The property of being at most countable is often referred to as being countable because the question of interest is normally whether one can list all elements of the set, designating a first, second, third etc. in such a way as to give each element of the set a natural number. The possibility that a single element of the set may be counted more than once is often not important.

Theorem 2.2.7 *If X and Y are both at most countable, then $X \times Y$ is also at most countable. If either X or Y is countable, then $X \times Y$ is also countable.*

Proof: It is given that there exists a mapping $\eta : \mathbb{N} \to X$ which is onto. Define $\eta(i) \equiv x_i$ and consider X as the set $\{x_1, x_2, x_3, \cdots\}$. Similarly, consider Y as the set $\{y_1, y_2, y_3, \cdots\}$. It follows the elements of $X \times Y$ are included in the following rectangular array.

$$\begin{array}{llll} (x_1,y_1) & (x_1,y_2) & (x_1,y_3) & \cdots \quad \leftarrow \text{Those which have } x_1 \text{ in first slot.} \\ (x_2,y_1) & (x_2,y_2) & (x_2,y_3) & \cdots \quad \leftarrow \text{Those which have } x_2 \text{ in first slot.} \\ (x_3,y_1) & (x_3,y_2) & (x_3,y_3) & \cdots \quad \leftarrow \text{Those which have } x_3 \text{ in first slot.} \\ \vdots & \vdots & \vdots & \qquad\qquad \vdots \end{array}$$

Follow a path through this array as follows.

$$\begin{array}{lllll} (x_1,y_1) & \to & (x_1,y_2) & & (x_1,y_3) \quad \to \\ & \swarrow & & \nearrow & \\ (x_2,y_1) & & (x_2,y_2) & & \\ \downarrow & \nearrow & & & \\ (x_3,y_1) & & & & \end{array}$$

Thus the first element of $X \times Y$ is (x_1,y_1), the second element of $X \times Y$ is (x_1,y_2), the third element of $X \times Y$ is (x_2,y_1) etc. This assigns a number from $\mathbb{N}$ to each element of $X \times Y$. Thus $X \times Y$ is at most countable.

It remains to show the last claim. Suppose without loss of generality that X is countable. Then there exists $\alpha : \mathbb{N} \to X$ which is one to one and onto. Let $\beta : X \times Y \to \mathbb{N}$ be defined by $\beta((x,y)) \equiv \alpha^{-1}(x)$. Thus β is onto $\mathbb{N}$. By the first part there exists a function from $\mathbb{N}$ onto $X \times Y$. Therefore, by Corollary 2.2.5, there exists a one to one and onto mapping from $X \times Y$ to $\mathbb{N}$. ■

Theorem 2.2.8 *If X and Y are at most countable, then $X \cup Y$ is at most countable. If either X or Y are countable, then $X \cup Y$ is countable.*

Proof: As in the preceding theorem,

$$X = \{x_1, x_2, x_3, \cdots\}$$

and

$$Y = \{y_1, y_2, y_3, \cdots\}.$$

Consider the following array consisting of $X \cup Y$ and path through it.

$$\begin{array}{llllll} x_1 & \to & x_2 & & x_3 & \to \\ & \swarrow & & \nearrow & & \\ y_1 & \to & y_2 & & & \end{array}$$

Thus the first element of $X \cup Y$ is x_1, the second is x_2 the third is y_1 the fourth is y_2 etc.

Consider the second claim. By the first part, there is a map from $\mathbb{N}$ onto $X \times Y$. Suppose without loss of generality that X is countable and $\alpha : \mathbb{N} \to X$ is one to one and onto. Then define $\beta(y) \equiv 1$, for all $y \in Y$,and $\beta(x) \equiv \alpha^{-1}(x)$. Thus, β maps $X \times Y$ onto $\mathbb{N}$ and this shows there exist two onto maps, one mapping $X \cup Y$ onto $\mathbb{N}$ and the other mapping $\mathbb{N}$ onto $X \cup Y$. Then Corollary 2.2.5 yields the conclusion. ■

Note that by induction this shows that if you have any finite set whose elements are countable sets, then the union of these is countable.

2.3 Equivalence Relations

There are many ways to compare elements of a set other than to say two elements are equal or the same. For example, in the set of people let two people be equivalent if they have the same weight. This would not be saying they were the same person, just that they weighed the same. Often such relations involve considering one characteristic of the elements of a set and then saying the two elements are equivalent if they are the same as far as the given characteristic is concerned.

Definition 2.3.1 *Let S be a set. $\sim$ is an equivalence relation on S if it satisfies the following axioms.*

1. $x \sim x$ *for all* $x \in S$. *(Reflexive)*
2. *If* $x \sim y$ *then* $y \sim x$. *(Symmetric)*
3. *If* $x \sim y$ *and* $y \sim z$, *then* $x \sim z$. *(Transitive)*

Definition 2.3.2 $[x]$ *denotes the set of all elements of S which are equivalent to x and $[x]$ is called the equivalence class determined by x or just the equivalence class of x.*

With the above definition one can prove the following simple theorem.

Theorem 2.3.3 *Let $\sim$ be an equivalence relation defined on a set, S and let $\mathscr{H}$ denote the set of equivalence classes. Then if $[x]$ and $[y]$ are two of these equivalence classes, either $x \sim y$ and $[x] = [y]$ or it is not true that $x \sim y$ and $[x] \cap [y] = \emptyset$.*

2.4 sup and inf

It is assumed in all that is done that $\mathbb{R}$ is complete. There are two ways to describe completeness of $\mathbb{R}$. One is to say that every bounded set has a least upper bound and a greatest lower bound. The other is to say that every Cauchy sequence converges. These two equivalent notions of completeness will be taken as given. Cauchy sequences are discussed a little later.

The symbol, $\mathbb{F}$ will mean either $\mathbb{R}$ or $\mathbb{C}$. The symbol $[-\infty, \infty]$ will mean all real numbers along with $+\infty$ and $-\infty$ which are points which we pretend are at the right and left ends of the real line respectively. The inclusion of these make believe points makes the statement of certain theorems less trouble.

Definition 2.4.1 *For $A \subseteq [-\infty, \infty], A \neq \emptyset$ $\sup A$ is defined as the least upper bound in case A is bounded above by a real number and equals ∞ if A is not bounded above. Similarly* $\inf A$ *is defined to equal the greatest lower bound in case A is bounded below by a real number and equals $-\infty$ in case A is not bounded below.*

Lemma 2.4.2 *If $\{A_n\}$ is an increasing sequence in $[-\infty, \infty]$, then*

$$\sup \{A_n : n \in \mathbb{N}\} = \lim_{n \to \infty} A_n.$$

Similarly, if $\{A_n\}$ is decreasing, then

$$\inf \{A_n : n \in \mathbb{N}\} = \lim_{n \to \infty} A_n.$$

Proof: Let $\sup(\{A_n : n \in \mathbb{N}\}) = r$. In the first case, suppose $r < \infty$. Then letting $\varepsilon > 0$ be given, there exists n such that $A_n \in (r-\varepsilon, r]$. Since $\{A_n\}$ is increasing, it follows if $m > n$, then $r - \varepsilon < A_n \leq A_m \leq r$ and so $\lim_{n\to\infty} A_n = r$ as claimed. In the case where $r = \infty$, then if a is a real number, there exists n such that $A_n > a$. Since $\{A_k\}$ is increasing, it follows that if $m > n$, $A_m > a$. But this is what is meant by $\lim_{n\to\infty} A_n = \infty$. The other case is that $r = -\infty$. But in this case, $A_n = -\infty$ for all n and so $\lim_{n\to\infty} A_n = -\infty$. The case where A_n is decreasing is entirely similar. ■

2.5 Double Series

Double series are of the form $\sum_{k=m}^{\infty}\sum_{j=m}^{\infty} a_{jk} \equiv \sum_{k=m}^{\infty}\left(\sum_{j=m}^{\infty} a_{jk}\right)$. In other words, first sum on j yielding something which depends on k and then sum these. The major consideration for these double series is the question of when

$$\sum_{k=m}^{\infty}\sum_{j=m}^{\infty} a_{jk} = \sum_{j=m}^{\infty}\sum_{k=m}^{\infty} a_{jk}$$

In other words, when does it make no difference which subscript is summed over first? In the case of finite sums there is no issue here. You can always write $\sum_{k=m}^{M}\sum_{j=m}^{N} a_{jk} = \sum_{j=m}^{N}\sum_{k=m}^{M} a_{jk}$ because addition is commutative. However, there are limits involved with infinite sums and the interchange in order of summation involves taking limits in a different order. Therefore, it is not always true that it is permissible to interchange the two sums. A general rule of thumb is this: If something involves changing the order in which two limits are taken, you may not do it without agonizing over the question. In general, limits foul up algebra and also introduce things which are counter intuitive. Here is an example. This example is a little technical. It is placed here just to prove conclusively there is a question which needs to be considered.

Example 2.5.1 *Consider the following picture which depicts some of the ordered pairs (m,n) where m,n are positive integers.*

$$\begin{array}{ccccccc} & & & \vdots & & & \\ 0 & 0 & c & 0 & -c & & \\ 0 & c & 0 & -c & 0 & & \\ b & 0 & -c & 0 & 0 & \cdots & \\ 0 & a & 0 & 0 & 0 & & \end{array}$$

The a,b,c are the values of a_{mn}. Thus $a_{nn} = 0$ for all $n \geq 1$, $a_{21} = a, a_{12} = b, a_{m(m+1)} = -c$ whenever $m > 1$, and $a_{m(m-1)} = c$ whenever $m > 2$. The numbers next to the point are the values of a_{mn}. You see $a_{nn} = 0$ for all n, $a_{21} = a, a_{12} = b, a_{mn} = c$ for (m,n) on the line $y = 1 + x$ whenever $m > 1$, and $a_{mn} = -c$ for all (m,n) on the line $y = x - 1$ whenever $m > 2$.

Then $\sum_{m=1}^{\infty} a_{mn} = a$ if $n = 1$, $\sum_{m=1}^{\infty} a_{mn} = b - c$ if $n = 2$ and if $n > 2, \sum_{m=1}^{\infty} a_{mn} = 0$. Therefore,

$$\sum_{n=1}^{\infty}\sum_{m=1}^{\infty} a_{mn} = a + b - c.$$

Next observe that $\sum_{n=1}^{\infty} a_{mn} = b$ if $m = 1, \sum_{n=1}^{\infty} a_{mn} = a + c$ if $m = 2$, and $\sum_{n=1}^{\infty} a_{mn} = 0$ if $m > 2$. Therefore,

$$\sum_{m=1}^{\infty} \sum_{n=1}^{\infty} a_{mn} = b + a + c$$

and so the two sums are different. Moreover, you can see that by assigning different values of a, b, and c, you can get an example for any two different numbers desired.

It turns out that if $a_{ij} \geq 0$ for all i, j, then you can always interchange the order of summation. This is shown next and is based on the following lemma. First, some notation should be discussed.

Definition 2.5.2 *Let $f(a,b) \in [-\infty, \infty]$ for $a \in A$ and $b \in B$ where A, B are sets which means that $f(a,b)$ is either a number, ∞, or $-\infty$. The symbol, $+\infty$ is interpreted as a point out at the end of the number line which is larger than every real number. Of course there is no such number. That is why it is called ∞. The symbol, $-\infty$ is interpreted similarly. Then $\sup_{a \in A} f(a,b)$ means $\sup(S_b)$ where $S_b \equiv \{f(a,b) : a \in A\}$.*

Unlike limits, you can take the sup in different orders.

Lemma 2.5.3 *Let $f(a,b) \in [-\infty, \infty]$ for $a \in A$ and $b \in B$ where A, B are sets. Then*

$$\sup_{a \in A} \sup_{b \in B} f(a,b) = \sup_{b \in B} \sup_{a \in A} f(a,b).$$

Proof: Note that for all a, b, $f(a,b) \leq \sup_{b \in B} \sup_{a \in A} f(a,b)$ and therefore, for all a, $\sup_{b \in B} f(a,b) \leq \sup_{b \in B} \sup_{a \in A} f(a,b)$. Therefore,

$$\sup_{a \in A} \sup_{b \in B} f(a,b) \leq \sup_{b \in B} \sup_{a \in A} f(a,b).$$

Repeat the same argument interchanging a and b, to get the conclusion of the lemma. ■

Theorem 2.5.4 *Let $a_{ij} \geq 0$. Then $\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} a_{ij} = \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} a_{ij}$.*

Proof: First note there is no trouble in defining these sums because the a_{ij} are all nonnegative. If a sum diverges, it only diverges to ∞ and so ∞ is the value of the sum. Next note that

$$\sum_{j=r}^{\infty} \sum_{i=r}^{\infty} a_{ij} \geq \sup_n \sum_{j=r}^{\infty} \sum_{i=r}^{n} a_{ij}$$

because for all $j, \sum_{i=r}^{\infty} a_{ij} \geq \sum_{i=r}^{n} a_{ij}$.Therefore,

$$\sum_{j=r}^{\infty} \sum_{i=r}^{\infty} a_{ij} \geq \sup_n \sum_{j=r}^{\infty} \sum_{i=r}^{n} a_{ij} = \sup_n \lim_{m \to \infty} \sum_{j=r}^{m} \sum_{i=r}^{n} a_{ij}$$

$$= \sup_n \lim_{m \to \infty} \sum_{i=r}^{n} \sum_{j=r}^{m} a_{ij} = \sup_n \sum_{i=r}^{n} \lim_{m \to \infty} \sum_{j=r}^{m} a_{ij}$$

$$= \sup_n \sum_{i=r}^{n} \sum_{j=r}^{\infty} a_{ij} = \lim_{n \to \infty} \sum_{i=r}^{n} \sum_{j=r}^{\infty} a_{ij} = \sum_{i=r}^{\infty} \sum_{j=r}^{\infty} a_{ij}$$

Interchanging the i and j in the above argument proves the theorem. ■

2.6 lim sup and lim inf

Sometimes the limit of a sequence does not exist. For example, if $a_n = (-1)^n$, then $\lim_{n\to\infty} a_n$ does not exist. This is because the terms of the sequence are a distance of 1 apart. Therefore there can't exist a single number such that all the terms of the sequence are ultimately within $1/4$ of that number. The nice thing about lim sup and lim inf is that they always exist. First here is a simple lemma and definition. First review the definition of inf and sup on Page 63 along with the simple properties of these things.

Definition 2.6.1 *Denote by $[-\infty,\infty]$ the real line along with symbols ∞ and $-\infty$. It is understood that ∞ is larger than every real number and $-\infty$ is smaller than every real number. Then if $\{A_n\}$ is an increasing sequence of points of $[-\infty,\infty]$, $\lim_{n\to\infty} A_n$ equals ∞ if the only upper bound of the set $\{A_n\}$ is ∞. If $\{A_n\}$ is bounded above by a real number, then $\lim_{n\to\infty} A_n$ is defined in the usual way and equals the least upper bound of $\{A_n\}$. If $\{A_n\}$ is a decreasing sequence of points of $[-\infty,\infty]$, $\lim_{n\to\infty} A_n$ equals $-\infty$ if the only lower bound of the sequence $\{A_n\}$ is $-\infty$. If $\{A_n\}$ is bounded below by a real number, then $\lim_{n\to\infty} A_n$ is defined in the usual way and equals the greatest lower bound of $\{A_n\}$. More simply, if $\{A_n\}$ is increasing, $\lim_{n\to\infty} A_n \equiv \sup\{A_n\}$ and if $\{A_n\}$ is decreasing then $\lim_{n\to\infty} A_n \equiv \inf\{A_n\}$.*

Lemma 2.6.2 *Let $\{a_n\}$ be a sequence of real numbers and let $U_n \equiv \sup\{a_k : k \geq n\}$. Then $\{U_n\}$ is a decreasing sequence. Also if $L_n \equiv \inf\{a_k : k \geq n\}$, then $\{L_n\}$ is an increasing sequence. Therefore, $\lim_{n\to\infty} L_n$ and $\lim_{n\to\infty} U_n$ both exist.*

Proof: Let W_n be an upper bound for $\{a_k : k \geq n\}$. Then since these sets are getting smaller, it follows that for $m < n$, W_m is an upper bound for $\{a_k : k \geq n\}$. In particular if $W_m = U_m$, then U_m is an upper bound for $\{a_k : k \geq n\}$ and so U_m is at least as large as U_n, the least upper bound for $\{a_k : k \geq n\}$. The claim that $\{L_n\}$ is decreasing is similar. ■

From the lemma, the following definition makes sense.

Definition 2.6.3 *Let $\{a_n\}$ be any sequence of points of $[-\infty,\infty]$*

$$\limsup_{n\to\infty} a_n \equiv \lim_{n\to\infty} \sup\{a_k : k \geq n\}$$

$$\liminf_{n\to\infty} a_n \equiv \lim_{n\to\infty} \inf\{a_k : k \geq n\}.$$

Theorem 2.6.4 *Suppose $\{a_n\}$ is a sequence of real numbers and also that*

$$\limsup_{n\to\infty} a_n, \liminf_{n\to\infty} a_n$$

are both real numbers. Then $\lim_{n\to\infty} a_n$ exists if and only if the two numbers are equal and in this case,

$$\lim_{n\to\infty} a_n = \liminf_{n\to\infty} a_n = \limsup_{n\to\infty} a_n.$$

Proof: First note that $\sup\{a_k : k \geq n\} \geq \inf\{a_k : k \geq n\}$ and so,

$$\limsup_{n\to\infty} a_n \equiv \lim_{n\to\infty} \sup\{a_k : k \geq n\} \geq \lim_{n\to\infty} \inf\{a_k : k \geq n\} \equiv \liminf_{n\to\infty} a_n.$$

Suppose first that $\lim_{n\to\infty} a_n$ exists and is a real number a. Then from the definition of a limit, there exists N corresponding to $\varepsilon/6$ in the definition. Hence, if $m,n \geq N$, then

$$|a_n - a_m| \leq |a_n - a| + |a - a_n| < \frac{\varepsilon}{6} + \frac{\varepsilon}{6} = \frac{\varepsilon}{3}.$$

From the definition of $\sup\{a_k : k \geq N\}$, there exists $n_1 \geq N$ such that

$$\sup\{a_k : k \geq N\} \leq a_{n_1} + \varepsilon/3.$$

Similarly, there exists $n_2 \geq N$ such that

$$\inf\{a_k : k \geq N\} \geq a_{n_2} - \varepsilon/3.$$

It follows that

$$\sup\{a_k : k \geq N\} - \inf\{a_k : k \geq N\} \leq |a_{n_1} - a_{n_2}| + \frac{2\varepsilon}{3} < \varepsilon.$$

Since the sequence, $\{\sup\{a_k : k \geq N\}\}_{N=1}^{\infty}$ is decreasing and $\{\inf\{a_k : k \geq N\}\}_{N=1}^{\infty}$ is increasing, it follows that

$$0 \leq \lim_{N\to\infty} \sup\{a_k : k \geq N\} - \lim_{N\to\infty} \inf\{a_k : k \geq N\} \leq \varepsilon$$

Since ε is arbitrary, this shows

$$\lim_{N\to\infty} \sup\{a_k : k \geq N\} = \lim_{N\to\infty} \inf\{a_k : k \geq N\} \tag{2.1}$$

Next suppose 2.1 and both equal $a \in \mathbb{R}$. Then

$$\lim_{N\to\infty} (\sup\{a_k : k \geq N\} - \inf\{a_k : k \geq N\}) = 0$$

Since $\sup\{a_k : k \geq N\} \geq \inf\{a_k : k \geq N\}$ it follows that for every $\varepsilon > 0$, there exists N such that

$$\sup\{a_k : k \geq N\} - \inf\{a_k : k \geq N\} < \varepsilon,$$

and for every N,

$$\inf\{a_k : k \geq N\} \leq a \leq \sup\{a_k : k \geq N\}$$

Thus if $n \geq N$, $|a - a_n| < \varepsilon$ which implies that $\lim_{n\to\infty} a_n = a$. In case

$$a = \infty = \lim_{N\to\infty} \sup\{a_k : k \geq N\} = \lim_{N\to\infty} \inf\{a_k : k \geq N\}$$

then if $r \in \mathbb{R}$ is given, there exists N such that $\inf\{a_k : k \geq N\} > r$ which is to say that $\lim_{n\to\infty} a_n = \infty$. The case where $a = -\infty$ is similar except you use $\sup\{a_k : k \geq N\}$. ■

The significance of lim sup and lim inf, in addition to what was just discussed, is contained in the following theorem which follows quickly from the definition.

Theorem 2.6.5 *Suppose $\{a_n\}$ is a sequence of points of $[-\infty,\infty]$. Also define $\lambda = \limsup_{n\to\infty} a_n$. Then if $b > \lambda$, it follows there exists N such that whenever $n \geq N, a_n \leq b$.If $c < \lambda$, then $a_n > c$ for infinitely many values of n. Let $\gamma = \liminf_{n\to\infty} a_n$.Then if $d < \gamma$, it follows there exists N such that whenever $n \geq N, a_n \geq d$. If $e > \gamma$, it follows $a_n < e$ for infinitely many values of n.*

The proof of this theorem is left as an exercise for you. It follows directly from the definition and it is the sort of thing you must do yourself. Here is one other simple proposition.

Proposition 2.6.6 *Let* $\lim_{n\to\infty} a_n = a > 0$. *Then* $\limsup_{n\to\infty} a_n b_n = a \limsup_{n\to\infty} b_n$.

Proof: This follows from the definition. Let $\lambda_n = \sup\{a_k b_k : k \geq n\}$. For all n large enough, $a_n > a - \varepsilon$ where ε is small enough that $a - \varepsilon > 0$. Therefore,

$$\lambda_n \geq \sup\{b_k : k \geq n\}(a - \varepsilon)$$

for all n large enough. Then

$$\begin{aligned} \limsup_{n\to\infty} a_n b_n &= \lim_{n\to\infty} \lambda_n \equiv \limsup_{n\to\infty} a_n b_n \geq \lim_{n\to\infty} (\sup\{b_k : k \geq n\}(a - \varepsilon)) \\ &= (a - \varepsilon) \limsup_{n\to\infty} b_n \end{aligned}$$

Similar reasoning shows $\limsup_{n\to\infty} a_n b_n \leq (a + \varepsilon) \limsup_{n\to\infty} b_n$. Now since $\varepsilon > 0$ is arbitrary, the conclusion follows. ■

2.7 Nested Interval Lemma

The nested interval lemma is a simple and important lemma which is used later quite a bit.

Lemma 2.7.1 *Let* $[a_k, b_k] \supseteq [a_{k+1}, b_{k+1}]$ *for all* $k = 1, 2, 3, \cdots$. *Then there exists a point* p *in* $\cap_{k=1}^{\infty} [a_k, b_k]$. *If* $\lim_{k\to\infty} (b_k - a_k) = 0$, *then there is only one such point*

Proof: We note that for any $k, l, a_k \leq b_l$. Here is why. If $k \leq l$, then $a_k \leq a_l \leq b_l$. If $k > l$, then $b_l \geq b_k \geq a_k$. It follows that for each l, $\sup_k a_k \leq b_l$. Hence $\sup_k a_k$ is a lower bound to the set of all b_l and so it is no larger than the greatest lower bound. It follows that $\sup_k a_k \leq \inf_l b_l$. Pick $x \in [\sup_k a_k, \inf_l b_l]$. Then for every $k, a_k \leq x \leq b_k$. Hence $x \in \cap_{k=1}^{\infty} [a_k, b_k]$.

To see the last claim, if q is another point in all the intervals, then both p and q are in $[a_k, b_k]$ and so $|p - q| \leq (b_k - a_k) < \varepsilon$ if k is large enough. Since ε is arbitrary, $p = q$. ■

2.8 The Hausdorff Maximal Theorem

This major theorem, or something like it (Several equivalent statements are proved later.), is either absolutely essential or extremely convenient. First is the definition of what is meant by a partial order.

Definition 2.8.1 *A nonempty set* $\mathscr{F}$ *is called a partially ordered set if it has a partial order denoted by* $\prec$. *This means it satisfies the following. If* $x \prec y$ *and* $y \prec z$, *then* $x \prec z$. *Also* $x \prec x$. *It is like* $\subseteq$ *on the set of all subsets of a given set. It is not the case that given two elements of* $\mathscr{F}$ *that they are related. In other words, you cannot conclude that either* $x \prec y$ *or* $y \prec x$. *A chain, denoted by* $\mathscr{C} \subseteq \mathscr{F}$ *has the property that it is totally ordered meaning that if* $x, y \in \mathscr{C}$, *either* $x \prec y$ *or* $y \prec x$. *A maximal chain is a chain* $\mathscr{C}$ *which has the property that there is no strictly larger chain. In other words, if* $x \in \mathscr{F} \setminus \cup\mathscr{C}$, *then* $\mathscr{C} \cup \{x\}$ *is no longer a chain so* x *fails to be related to something in* $\mathscr{C}$.

Here is the Hausdorff maximal theorem. The proof is a proof by contradiction. We assume there is no maximal chain and then show this cannot happen. The axiom of choice is used in choosing the $x_{\mathscr{C}}$ right at the beginning of the argument.

Theorem 2.8.2 *Let $\mathscr{F}$ be a nonempty partially ordered set with order $\prec$. Then there exists a maximal chain.*

Proof: Suppose no chain is maximal. Then, from the axiom of choice, for each chain $\mathscr{C}$ there exists $x_{\mathscr{C}} \in \mathscr{F} \setminus \cup\mathscr{C}$ such that $\mathscr{C} \cup \{x_{\mathscr{C}}\}$ is a chain. Call two chains comparable if one is a subset of the other. Let $\mathscr{X}$ be the set of all chains. Thus the elements of $\mathscr{X}$ are chains and $\mathscr{X}$ is a subset of $\mathscr{P}(\mathscr{F})$. For $\mathscr{C}$ a chain, let $\theta\mathscr{C}$ denote $\mathscr{C} \cup \{x_{\mathscr{C}}\}$. Pick $x_0 \in \mathscr{F}$. A subset $\mathscr{Y}$ of $\mathscr{X}$ will be called a "tower" if $\theta\mathscr{C} \in \mathscr{Y}$ whenever $\mathscr{C} \in \mathscr{Y}$, $x_0 \in \cup\mathscr{C}$ for all $\mathscr{C} \in \mathscr{Y}$, and if $\mathscr{S}$ is a subset of $\mathscr{Y}$ such that all pairs of chains in $\mathscr{S}$ are comparable, then $\cup\mathscr{S}$ is in $\mathscr{Y}$. Note that $\mathscr{X}$ is a tower. Let $\mathscr{Y}_0$ be the intersection of all towers. Then $\mathscr{Y}_0$ is also a tower, the smallest one. This follows from the definition.

Claim 1: If $\mathscr{C}_0 \in \mathscr{Y}_0$ is comparable to every chain $\mathscr{C} \in \mathscr{Y}_0$, then if $\mathscr{C}_0 \subsetneq \mathscr{C}$, it must be the case that $\theta\mathscr{C}_0 \subseteq \mathscr{C}$. The symbol $\subsetneq$ indicates proper subset.

Proof of Claim 1: Consider $\mathscr{B} \equiv \left\{\mathscr{D} \in \mathscr{Y}_0 : \mathscr{D} \subseteq \mathscr{C}_0 \text{ or } x_{\mathscr{C}_0} \in \cup\mathscr{D}\right\}$. Let $\mathscr{Y}_1 \equiv \mathscr{Y}_0 \cap \mathscr{B}$. I want to argue that $\mathscr{Y}_1$ is a tower. Obviously all chains of $\mathscr{Y}_1$ contain x_0 in their unions. If $\mathscr{D} \in \mathscr{Y}_1$, is $\theta\mathscr{D} \in \mathscr{Y}_1$?

case 1: $\mathscr{D} \supsetneq \mathscr{C}_0$. Then $x_{\mathscr{C}_0} \in \cup\mathscr{D}$ and so $x_{\mathscr{C}_0} \in \cup\theta\mathscr{D}$ so $\theta\mathscr{D} \in \mathscr{Y}_1$.

case 2: $\mathscr{D} \subseteq \mathscr{C}_0$. Then if $\theta\mathscr{D} \supsetneq \mathscr{C}_0$, it follows that $\mathscr{D} \subseteq \mathscr{C}_0 \subsetneq \mathscr{D} \cup \{x_{\mathscr{D}}\}$. If $x \in \mathscr{C}_0 \setminus \mathscr{D}$ then $x = x_{\mathscr{D}}$. But then

$$\mathscr{C}_0 \subsetneq \mathscr{D} \cup \{x_{\mathscr{D}}\} \subseteq \mathscr{C}_0 \cup \{x_{\mathscr{D}}\} = \mathscr{C}_0$$

which is nonsense, and so $\mathscr{C}_0 = \mathscr{D}$ so $x_{\mathscr{D}} = x_{\mathscr{C}_0} \in \cup\theta\mathscr{C}_0 = \cup\theta\mathscr{D}$ and so $\theta\mathscr{D} \in \mathscr{Y}_1$. If $\theta\mathscr{D} \subseteq \mathscr{C}_0$ then right away $\theta\mathscr{D} \in \mathscr{B}$. Thus $\mathscr{B} = \mathscr{Y}_0$ because $\mathscr{Y}_1$ cannot be smaller than $\mathscr{Y}_0$. In particular, if $\mathscr{D} \supsetneq \mathscr{C}_0$, then $x_{\mathscr{C}_0} \in \cup\mathscr{D}$ or in other words, $\theta\mathscr{C}_0 \subseteq \mathscr{D}$.

Claim 2: Any two chains in $\mathscr{Y}_0$ are comparable so if $\mathscr{C} \subsetneq \mathscr{D}$, then $\theta\mathscr{C} \subseteq \mathscr{D}$.

Proof of Claim 2: Let $\mathscr{Y}_1$ consist of all chains of $\mathscr{Y}_0$ which are comparable to every chain of $\mathscr{Y}_0$. (Like $\mathscr{C}_0$ above.) I want to show that $\mathscr{Y}_1$ is a tower. Let $\mathscr{C} \in \mathscr{Y}_1$ and $\mathscr{D} \in \mathscr{Y}_0$. Since $\mathscr{C}$ is comparable to all chains in $\mathscr{Y}_0$, either $\mathscr{C} \subsetneq \mathscr{D}$ or $\mathscr{C} \supseteq \mathscr{D}$. I need to show that $\theta\mathscr{C}$ is comparable with $\mathscr{D}$. The second case is obvious so consider the first that $\mathscr{C} \subsetneq \mathscr{D}$. By Claim 1, $\theta\mathscr{C} \subseteq \mathscr{D}$. Since $\mathscr{D}$ is arbitrary, this shows that $\mathscr{Y}_1$ is a tower. Hence $\mathscr{Y}_1 = \mathscr{Y}_0$ because $\mathscr{Y}_0$ is as small as possible. It follows that every two chains in $\mathscr{Y}_0$ are comparable and so if $\mathscr{C} \subsetneq \mathscr{D}$, then $\theta\mathscr{C} \subseteq \mathscr{D}$.

Since every pair of chains in $\mathscr{Y}_0$ are comparable and $\mathscr{Y}_0$ is a tower, it follows that $\cup\mathscr{Y}_0 \in \mathscr{Y}_0$ so $\cup\mathscr{Y}_0$ is a chain. However, $\theta \cup \mathscr{Y}_0$ is a chain which properly contains $\cup\mathscr{Y}_0$ and since $\mathscr{Y}_0$ is a tower, $\theta \cup \mathscr{Y}_0 \in \mathscr{Y}_0$. Thus $\cup(\theta \cup \mathscr{Y}_0) \supsetneq \cup(\cup\mathscr{Y}_0) \supseteq \cup(\theta \cup \mathscr{Y}_0)$ which is a contradiction. Therefore, it is impossible to obtain the x_C described above for some chain $\mathscr{C}$ and so, this $\mathscr{C}$ is a maximal chain. ■

If X is a nonempty set, $\leq$ is an order on X if

$$\begin{gathered} x \leq x, \\ \text{either } x \leq y \text{ or } y \leq x \\ \text{if } x \leq y \text{ and } y \leq z \text{ then } x \leq z. \end{gathered}$$

and $\leq$ is a well order if $(X, \leq)$ if every nonempty subset of X has a smallest element. More precisely, if $S \neq \emptyset$ and $S \subseteq X$ then there exists an $x \in S$ such that $x \leq y$ for all $y \in S$. A familiar example of a well-ordered set is the natural numbers.

Lemma 2.8.3 *The Hausdorff maximal principle implies every nonempty set can be well-ordered.*

Proof: Let X be a nonempty set and let $a \in X$. Then $\{a\}$ is a well-ordered subset of X. Let

$$\mathscr{F} = \{S \subseteq X : \text{ there exists a well order for } S\}.$$

Thus $\mathscr{F} \neq \emptyset$. For $S_1,\ S_2 \in \mathscr{F}$, define $S_1 \prec S_2$ if $S_1 \subseteq S_2$ and there exists a well order for S_2, $\leq_2$ such that

$$(S_2, \leq_2) \text{ is well-ordered}$$

and if

$$y \in S_2 \setminus S_1 \text{ then } x \leq_2 y \text{ for all } x \in S_1,$$

and if $\leq_1$is the well order of S_1 then the two orders are consistent on S_1. Then observe that $\prec$ is a partial order on $\mathscr{F}$. By the Hausdorff maximal principle, let $\mathscr{C}$ be a maximal chain in $\mathscr{F}$ and let $X_\infty \equiv \cup\mathscr{C}$. Define an order, $\leq$, on X_∞ as follows. If $x,\ y$ are elements of X_∞, pick $S \in \mathscr{C}$ such that $x,\ y$ are both in S. Then if $\leq_S$ is the order on S, let $x \leq y$ if and only if $x \leq_S y$. This definition is well defined because of the definition of the order, $\prec$. Now let U be any nonempty subset of X_∞. Then $S \cap U \neq \emptyset$ for some $S \in \mathscr{C}$. Because of the definition of $\leq$, if $y \in S_2 \setminus S_1,\ S_i \in \mathscr{C}$, then $x \leq y$ for all $x \in S_1$. Thus, if $y \in X_\infty \setminus S$ then $x \leq y$ for all $x \in S$ and so the smallest element of $S \cap U$ exists and is the smallest element in U. Therefore X_∞ is well-ordered. Now suppose there exists $z \in X \setminus X_\infty$. Define the following order, $\leq_1$, on $X_\infty \cup \{z\}$.

$$x \leq_1 y \text{ if and only if } x \leq y \text{ whenever } x, y \in X_\infty$$

$$x \leq_1 z \text{ whenever } x \in X_\infty.$$

Then let

$$\widetilde{\mathscr{C}} = \{S \in \mathscr{C} \text{ or } X_\infty \cup \{z\}\}.$$

Then $\widetilde{\mathscr{C}}$ is a strictly larger chain than $\mathscr{C}$ contradicting maximality of $\mathscr{C}$. Thus $X \setminus X_\infty = \emptyset$ and this shows X is well-ordered by $\leq$. ■

With these two lemmas the main result follows.

Theorem 2.8.4 *The following are equivalent.*

The axiom of choice

The Hausdorff maximal principle

The well-ordering principle.

Proof: It remains to show that the well-ordering principle implies the axiom of choice. Let I be a nonempty set and let X_i be a nonempty set for each $i \in I$. Let $X = \cup\{X_i : i \in I\}$ and well order X. Let $f(i)$ be the smallest element of X_i. Then $f \in \prod_{i \in I} X_i$. ■

The book by Hewitt and Stromberg [26] has more equivalences.

Chapter 3

Metric Spaces

3.1 Open and Closed Sets, Sequences, Limit Points

It is most efficient to discus things in terms of abstract metric spaces to begin with.

Definition 3.1.1 *A non empty set X is called a metric space if there is a function $d : X \times X \to [0,\infty)$ which satisfies the following axioms.*

1. $d(x,y) = d(y,x)$
2. $d(x,y) \geq 0$ *and equals 0 if and only if* $x = y$
3. $d(x,y) + d(y,z) \geq d(x,z)$

This function d is called the metric. We often refer to it as the distance also.

Definition 3.1.2 *An open ball, denoted as $B(x,r)$ is defined as follows.*

$$B(x,r) \equiv \{y : d(x,y) < r\}$$

A set U is said to be open if whenever $x \in U$, it follows that there is $r > 0$ such that $B(x,r) \subseteq U$. More generally, a point x is said to be an interior point of U if there exists such a ball. In words, an open set is one for which every point is an interior point.

For example, you could have X be a subset of $\mathbb{R}$ and $d(x,y) = |x-y|$.
Then the first thing to show is the following.

Proposition 3.1.3 *An open ball is an open set.*

Proof: Suppose $y \in B(x,r)$. We need to verify that y is an interior point of $B(x,r)$. Let $\delta = r - d(x,y)$. Then if $z \in B(y,\delta)$, it follows that

$$d(z,x) \leq d(z,y) + d(y,x) < \delta + d(y,x) = r - d(x,y) + d(y,x) = r$$

Thus $y \in B(y,\delta) \subseteq B(x,r)$. ■

Definition 3.1.4 *Let S be a nonempty subset of a metric space. Then p is a limit point (accumulation point) of S if for every $r > 0$ there exists a point different than p in $B(p,r) \cap S$. Sometimes people denote the set of limit points as S'.*

The following proposition is fairly obvious from the above definition and will be used whenever convenient. It is equivalent to the above definition and so it can take the place of the above definition if desired.

Proposition 3.1.5 *A point x is a limit point of the nonempty set A if and only if every $B(x,r)$ contains infinitely many points of A.*

Proof: $\Leftarrow$ is obvious. Consider $\Rightarrow$. Let x be a limit point. Let $r_1 = 1$. Then $B(x, r_1)$ contains $a_1 \neq x$. If $\{a_1, \cdots, a_n\}$ have been chosen none equal to x and with no repeats in the list, let $0 < r_n < \min\left(\frac{1}{n}, \min\{d(a_i, x), i = 1, 2, \cdots n\}\right)$. Then let $a_{n+1} \in B(x, r_n)$. Thus every $B(x, r)$ contains $B(x, r_n)$ for all n large enough and hence it contains a_k for $k \geq n$ where the a_k are distinct, none equal to x. ■

A related idea is the notion of the limit of a sequence. Recall that a sequence is really just a mapping from $\mathbb{N}$ to X. We write them as $\{x_n\}$ or $\{x_n\}_{n=1}^{\infty}$ if we want to emphasize the values of n. Then the following definition is what it means for a sequence to converge.

Definition 3.1.6 *We say that $x = \lim_{n\to\infty} x_n$ when for every $\varepsilon > 0$ there exists N such that if $n \geq N$, then*

$$d(x, x_n) < \varepsilon$$

Often we write $x_n \to x$ for short. This is equivalent to saying

$$\lim_{n\to\infty} d(x, x_n) = 0.$$

Proposition 3.1.7 *The limit is well defined. That is, if x, x' are both limits of a sequence, then $x = x'$.*

Proof: From the definition, there exist N, N' such that if $n \geq N$, then $d(x, x_n) < \varepsilon/2$ and if $n \geq N'$, then $d(x, x_n) < \varepsilon/2$. Then let $M \geq \max(N, N')$. Let $n > M$. Then

$$d(x, x') \leq d(x, x_n) + d(x_n, x') < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

Since ε is arbitrary, this shows that $x = x'$ because $d(x, x') = 0$. ■

Next there is an important theorem about limit points and convergent sequences.

Theorem 3.1.8 *Let $S \neq \emptyset$. Then p is a limit point of S if and only if there exists a sequence of distinct points of S, $\{x_n\}$ none of which equal p such that $\lim_{n\to\infty} x_n = p$.*

Proof: $\Longrightarrow$ Suppose p is a limit point. Why does there exist the promissed convergent sequence? Let $x_1 \in B(p, 1) \cap S$ such that $x_1 \neq p$. If $x_1, \cdots, x_n$ have been chosen, let $x_{n+1} \neq p$ be in $B(p, \delta_{n+1}) \cap S$ where

$$\delta_{n+1} = \min\left\{\frac{1}{n+1}, d(x_i, p), i = 1, 2, \cdots, n\right\}.$$

Then this constructs the necessary convergent sequence.

$\Longleftarrow$ Conversely, if such a sequence $\{x_n\}$ exists, then for every $r > 0$, $B(p, r)$ contains $x_n \in S$ for all n large enough. Hence, p is a limit point because none of these x_n are equal to p. ■

Definition 3.1.9 *A set H is closed means H^C is open.*

Note that this says that the complement of an open set is closed. If V is open, then the complement of its complement is itself. Thus $\left(V^C\right)^C = V$ an open set. Hence V^C is closed.

Then the following theorem gives the relationship between closed sets and limit points.

Theorem 3.1.10 *A set H is closed if and only if it contains all of its limit points.*

Proof: $\Longrightarrow$ Let H be closed and let p be a limit point. We need to verify that $p \in H$. If it is not, then since H is closed, its complement is open and so there exists $\delta > 0$ such that $B(p,\delta) \cap H = \emptyset$. However, this prevents p from being a limit point.

$\Longleftarrow$ Next suppose H has all of its limit points. Why is H^C open? If $p \in H^C$ then it is not a limit point and so there exists $\delta > 0$ such that $B(p,\delta)$ has no points of H. In other words, H^C is open. Hence H is closed. ■

Corollary 3.1.11 *A set H is closed if and only if whenever $\{h_n\}$ is a sequence of points of H which converges to a point x, it follows that $x \in H$.*

Proof: $\Longrightarrow$ Suppose H is closed and $h_n \to x$. If $x \in H$ there is nothing left to show. If $x \notin H$, then from the definition of limit, it is a limit point of H because none of the h_n are equal to x. Hence $x \in H$ after all.

$\Longleftarrow$ Suppose the limit condition holds, why is H closed? Let $x \in H'$ the set of limit points of H. By Theorem 3.1.8 there exists a sequence of points of H, $\{h_n\}$ such that $h_n \to x$. Then by assumption, $x \in H$. Thus H contains all of its limit points and so it is closed by Theorem 3.1.10. ■

Next is the important concept of a subsequence.

Definition 3.1.12 *Let $\{x_n\}_{n=1}^{\infty}$ be a sequence. Then if $n_1 < n_2 < \cdots$ is a strictly increasing sequence of indices, we say $\{x_{n_k}\}_{k=1}^{\infty}$ is a subsequence of $\{x_n\}_{n=1}^{\infty}$.*

The really important thing about subsequences is that they preserve convergence.

Theorem 3.1.13 *Let $\{x_{n_k}\}$ be a subsequence of a convergent sequence $\{x_n\}$ where $x_n \to x$. Then $\lim_{k\to\infty} x_{n_k} = x$ also.*

Proof: Let $\varepsilon > 0$ be given. Then there exists N such that $d(x_n,x) < \varepsilon$ if $n \geq N$. It follows that if $k \geq N$, then $n_k \geq N$ and so $d(x_{n_k},x) < \varepsilon$ if $k \geq N$. This is what it means to say $\lim_{k\to\infty} x_{n_k} = x$. ■

3.2 Cauchy Sequences, Completeness

Of course it does not go the other way. For example, you could let $x_n = (-1)^n$ and it has a convergent subsequence but fails to converge. Here $d(x,y) = |x-y|$ and the metric space is just $\mathbb{R}$.

However, there is a kind of sequence for which it does go the other way. This is called a Cauchy sequence.

Definition 3.2.1 *$\{x_n\}$ is called a Cauchy sequence if for every $\varepsilon > 0$ there exists N such that if $m,n \geq N$, then $d(x_n,x_m) < \varepsilon$.*

Now the major theorem about this is the following.

Theorem 3.2.2 *Let $\{x_n\}$ be a Cauchy sequence. Then it converges if and only if any subsequence converges.*

Proof: $\Longrightarrow$ This was just done above. $\Longleftarrow$ Suppose now that $\{x_n\}$ is a Cauchy sequence and $\lim_{k\to\infty} x_{n_k} = x$. Then there exists N_1 such that if $k > N_1$, then $d(x_{n_k},x) < \varepsilon/2$. From the definition of what it means to be Cauchy, there exists N_2 such that if $m,n \geq N_2$, then $d(x_m,x_n) < \varepsilon/2$. Let $N \geq \max(N_1,N_2)$. Then if $k \geq N$, then $n_k \geq N$ and so $d(x,x_k) \leq d(x,x_{n_k}) + d(x_{n_k},x_k) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$. It follows from the definition that $\lim_{k\to\infty} x_k = x$. ■

Definition 3.2.3 *A metric space is said to be* ***complete*** *if every Cauchy sequence converges.*

There certainly are metric spaces which are not complete. For example, if you consider $\mathbb{Q}$ with $d(x,y) \equiv |x-y|$, this will not be complete because you can get a sequence which is obtained as x_n defined as the n decimal place description of $\sqrt{2}$. However, if a sequence converges, then it must be Cauchy.

Lemma 3.2.4 *If $x_n \to x$, then $\{x_n\}$ is a Cauchy sequence.*

Proof: Let $\varepsilon > 0$. Then there exists n_ε such that if $m \geq n_\varepsilon$, then $d(x,x_m) < \varepsilon/2$. If $m,k \geq n_\varepsilon$, then by the triangle inequality, $d(x_m,x_k) \leq d(x_m,x) + d(x,x_k) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$ showing that the convergent sequence is indeed a Cauchy sequence as claimed. ■

Another nice thing to note is this.

Proposition 3.2.5 *If $\{x_n\}$ is a sequence and if p is a limit point of the set $S = \cup_{n=1}^{\infty}\{x_n\}$, then there is a subsequence $\{x_{n_k}\}$ such that $\lim_{k\to\infty} x_{n_k} = x$.*

Proof: By Theorem 3.1.8, there exists a sequence of distinct points of S denoted as $\{y_k\}$ such that none of them equal p and $\lim_{k\to\infty} y_k = p$. Thus $B(p,r)$ contains infinitely many different points of the set D, this for every r. Let $x_{n_1} \in B(p,1)$ where n_1 is the first index such that $x_{n_1} \in B(p,1)$. Suppose $x_{n_1}, \cdots, x_{n_k}$ have been chosen, the n_i increasing and let $1 > \delta_1 > \delta_2 > \cdots > \delta_k$ where $x_{n_i} \in B(p,\delta_i)$. Then let

$$\delta_{k+1} < \min\left\{\frac{1}{2^{k+1}}, d\left(p,x_{n_j}\right), \delta_j, j = 1,2\cdots,k\right\}$$

Let $x_{n_{k+1}} \in B(p,\delta_{k+1})$ where n_{k+1} is the first index such that $x_{n_{k+1}}$ is contained $B(p,\delta_{k+1})$. Then $\lim_{k\to\infty} x_{n_k} = p$. ■

Another useful result is the following.

Lemma 3.2.6 *Suppose $x_n \to x$ and $y_n \to y$. Then $d(x_n,y_n) \to d(x,y)$.*

Proof: Consider the following.

$$d(x,y) \leq d(x,x_n) + d(x_n,y) \leq d(x,x_n) + d(x_n,y_n) + d(y_n,y)$$

so $d(x,y) - d(x_n,y_n) \leq d(x,x_n) + d(y_n,y)$. Similar reasoning to what was just used shows that $d(x_n,y_n) - d(x,y) \leq d(x,x_n) + d(y_n,y)$, so $|d(x_n,y_n) - d(x,y)| \leq d(x,x_n) + d(y_n,y)$ and the right side converges to 0 as $n \to \infty$. ■

3.3 Closure of a Set

Next is the topic of the closure of a set.

Definition 3.3.1 *Let A be a nonempty subset of (X,d) a metric space. Then $\overline{A}$ is defined to be the intersection of all closed sets which contain A. Note the whole space, X is one such closed set which contains A. The whole space X is closed because its complement is open, its complement being $\emptyset$. It is certainly true that every point of the empty set is an interior point because there are no points of $\emptyset$.*

Lemma 3.3.2 *Let A be a nonempty set in (X,d). Then $\overline{A}$ is a closed set and $\overline{A} = A \cup A'$ where A' denotes the set of limit points of A.*

Proof: First of all, denote by $\mathscr{C}$ the set of closed sets which contain A. Then $\overline{A} = \cap\mathscr{C}$ and this will be closed if its complement is open. However, $\overline{A}^C = \cup\left\{H^C : H \in \mathscr{C}\right\}$. Each H^C is open and so the union of all these open sets must also be open. This is because if x is in this union, then it is in at least one of them. Hence it is an interior point of that one. But this implies it is an interior point of the union of them all which is an even larger set. Thus $\overline{A}$ is closed.

The interesting part is the next claim. First note that from the definition, $A \subseteq \overline{A}$ so if $x \in A$, then $x \in \overline{A}$. Now consider $y \in A'$ but $y \notin A$. If $y \notin \overline{A}$, a closed set, then there exists $B(y,r) \subseteq \overline{A}^C$. Thus y cannot be a limit point of A, a contradiction. Therefore, $A \cup A' \subseteq \overline{A}$.

Next suppose $x \in \overline{A}$ and suppose $x \notin A$. Then if $B(x,r)$ contains no points of A different than x, since x itself is not in A, it would follow that $B(x,r) \cap A = \emptyset$ and so recalling that open balls are open, $B(x,r)^C$ is a closed set containing A so from the definition, it also contains $\overline{A}$ which is contrary to the assertion that $x \in \overline{A}$. Hence if $x \notin A$, then $x \in A'$ and so $A \cup A' \supseteq \overline{A}$ ■

3.4 Separable Metric Spaces

Definition 3.4.1 *A metric space is called separable if there exists a countable dense subset D. This means two things. First, D is countable, and second, that if x is any point and $r > 0$, then $B(x,r) \cap D \neq \emptyset$. A metric space is called completely separable if there exists a countable collection of nonempty open sets $\mathscr{B}$ such that every open set is the union of some subset of $\mathscr{B}$. This collection of open sets is called a countable basis.*

For those who like to fuss about empty sets, the empty set is open and it is indeed the union of a subset of $\mathscr{B}$ namely the empty subset.

Theorem 3.4.2 *A metric space is separable if and only if it is completely separable.*

Proof: $\Longleftarrow$ Let $\mathscr{B}$ be the special countable collection of open sets and for each $B \in \mathscr{B}$, let p_B be a point of B. Then let $\mathscr{P} \equiv \{p_B : B \in \mathscr{B}\}$. If $B(x,r)$ is any ball, then it is the union of sets of $\mathscr{B}$ and so there is a point of $\mathscr{P}$ in it. Since $\mathscr{B}$ is countable, so is $\mathscr{P}$.

$\Longrightarrow$ Let D be the countable dense set and let $\mathscr{B} \equiv \{B(d,r) : d \in D, r \in \mathbb{Q} \cap [0,\infty)\}$. Then $\mathscr{B}$ is countable because the Cartesian product of countable sets is countable. It suffices to show that every ball is the union of these sets. Let $B(x,R)$ be a ball. Let $y \in B(y,\delta) \subseteq B(x,R)$. Then there exists $d \in B\left(y, \frac{\delta}{10}\right)$. Let $\varepsilon \in \mathbb{Q}$ and $\frac{\delta}{10} < \varepsilon < \frac{\delta}{5}$. Then $y \in B(d,\varepsilon) \in \mathscr{B}$. Is $B(d,\varepsilon) \subseteq B(x,R)$? If so, then the desired result follows because this would show that every $y \in B(x,R)$ is contained in one of these sets of $\mathscr{B}$ which is contained in $B(x,R)$ showing that $B(x,R)$ is the union of sets of $\mathscr{B}$. Let $z \in B(d,\varepsilon) \subseteq B\left(d, \frac{\delta}{5}\right)$. Then

$$d(y,z) \leq d(y,d) + d(d,z) < \frac{\delta}{10} + \varepsilon < \frac{\delta}{10} + \frac{\delta}{5} < \delta$$

Hence $B(d,\varepsilon) \subseteq B(y,\delta) \subseteq B(x,r)$. Therefore, every ball is the union of sets of $\mathscr{B}$ and, since every open set is the union of balls, it follows that every open set is the union of sets of $\mathscr{B}$. ■

Corollary 3.4.3 *If (X,d) is a metric space and S is a nonempty subset of X, then S is also separable.*

Proof: Let $\mathscr{B}$ be a countable basis for (X,d). Say $\mathscr{B}_S$ be those sets of $\mathscr{B}$ which have nonempty intersections with S. By axiom of choice, there is a point in each of these intersections. The resulting countable selection of points must be dense in S. Indeed, if $x \in S$, then $B(x,r)$ is the union of sets of $\mathscr{B}$ and so some point just described is in $B(x,r)$. ■

Definition 3.4.4 *Let S be a nonempty set. Then a set of open sets $\mathscr{C}$ is called an **open cover of** S if $\cup\mathscr{C} \supseteq \mathscr{S}$. (It covers up the set S. Think lilly pads covering the surface of a pond.)*

One of the important properties possessed by separable metric spaces is the Lindeloff property.

Definition 3.4.5 *A metric space has the Lindeloff property if whenever $\mathscr{C}$ is an open cover of a set S, there exists a countable subset of $\mathscr{C}$ denoted here by $\mathscr{B}$ such that $\mathscr{B}$ is also an open cover of S.*

Theorem 3.4.6 *Every separable metric space has the Lindeloff property.*

Proof: Let $\mathscr{C}$ be an open cover of a set S. Let $\mathscr{B}$ be a countable basis. Such exists by Theorem 3.4.2. Let $\hat{\mathscr{B}}$ denote those sets of $\mathscr{B}$ which are contained in some set of $\mathscr{C}$. Thus $\hat{\mathscr{B}}$ is a countable open cover of S. Now for $B \in \mathscr{B}$, let U_B be a set of $\mathscr{C}$ which contains B. Letting $\widehat{\mathscr{C}}$ denote these sets U_B it follows that $\widehat{\mathscr{C}}$ is countable and is an open cover of S. ■

Definition 3.4.7 *A Polish space is a complete separable metric space. These things turn out to be very useful in probability theory and in other areas.*

3.5 Compact Sets

As usual, we are not worrying about empty sets. Fussing over these is usually a waste of time. Thus if a set is mentioned, the default is that it is nonempty.

Definition 3.5.1 *A metric space K is compact if whenever $\mathscr{C}$ is an open cover of K, meaning $K \subseteq \cup\mathscr{C}$, there exists a finite subset of $\mathscr{C}$ $\{U_1,\cdots,U_n\}$ such that $K \subseteq \cup_{k=1}^n U_k$. In words, every open cover admits a finite sub-cover.*

Directly from this definition is the following proposition.

Proposition 3.5.2 *If K is a closed, nonempty subset of a nonempty compact set H, then K is compact.*

Proof: Let $\mathscr{C}$ be an open cover for K. Then $\mathscr{C} \cup \{K^C\}$ is an open cover for H. Thus there are finitely many sets from this last collection of open sets, $U_1,\cdots,U_m$ which covers H. Include only those which are in $\mathscr{C}$. These cover K because K^C covers no points of K. ■

This is the real definition given above. However, in metric spaces, it is equivalent to another definition called sequentially compact.

Definition 3.5.3 *A metric space K is sequentially compact means that whenever $\{x_n\} \subseteq K$, there exists a subsequence $\{x_{n_k}\}$ such that $\lim_{k\to\infty} x_{n_k} = x \in K$ for some point x. In words, every sequence has a subsequence which converges to a point in the set.*

There is a fundamental property possessed by a sequentially compact set in a metric space which is described in the following proposition. The special number described is called a Lebesgue number.

Proposition 3.5.4 *Let K be a sequentially compact set in a metric space and let $\mathscr{C}$ be an open cover of K. Then there exists a number $\delta > 0$ such that whenever $x \in K$, it follows that $B(x,\delta)$ is contained in some set of $\mathscr{C}$.*

Proof: If $\mathscr{C}$ is an open cover of K and has no Lebesgue number, then for each $n \in \mathbb{N}$, $\frac{1}{n}$ is not a Lebesgue number. Hence there exists $x_n \in K$ such that $B\left(x_n, \frac{1}{n}\right)$ is not contained in any set of $\mathscr{C}$. By sequential compactness, there is a subsequence $\{x_{n_k}\}$ such that $x_{n_k} \to x \in K$. Now there is $r > 0$ such that $B(x,r) \subseteq U \in \mathscr{C}$. Let k be large enough that $\frac{1}{n_k} < \frac{r}{2}$ and also large enough that $x_{n_k} \in B\left(x, \frac{r}{2}\right)$. Then $B\left(x_{n_k}, \frac{1}{n_k}\right) \subseteq B\left(x_{n_k}, \frac{r}{2}\right) \subseteq B(x,r)$ contrary to the requirement that $B\left(x_{n_k}, \frac{1}{n_k}\right)$ is not contained in any set of $\mathscr{C}$. ■

In any metric space, these two definitions of compactness are equivalent.

Theorem 3.5.5 *Let K be a nonempty subset of a metric space (X,d). Then it is compact if and only if it is sequentially compact.*

Proof: $\Leftarrow$ Suppose K is sequentially compact. Let $\mathscr{C}$ be an open cover of K. By Proposition 3.5.4 there is a Lebesgue number $\delta > 0$. Let $x_1 \in K$. If $B(x_1,\delta)$ covers K, then pick a set of $\mathscr{C}$ containing this ball and this set will be a finite subset of $\mathscr{C}$ which covers K. If $B(x_1,\delta)$ does not cover K, let $x_2 \notin B(x_1,\delta)$. Continue this way obtaining x_k such that $d(x_k,x_j) \geq \delta$ whenever $k \neq j$. Thus eventually $\{B(x_i,\delta)\}_{i=1}^n$ must cover K because if not, you could get a sequence $\{x_k\}$ which has every pair of points further apart than δ and hence it has no Cauchy subsequence. Therefore, by Lemma 3.2.4, it would have no convergent subsequence. This would contradict K is sequentially compact. Now let $U_i \in \mathscr{C}$ with $U_i \supseteq B(x_i,\delta)$. Then $\cup_{i=1}^n U_i \supseteq K$.

$\Rightarrow$ Now suppose K is compact. If it is not sequentially compact, then there exists a sequence $\{x_n\}$ which has no convergent subsequence to a point of K. In particular, no point of this sequence is repeated infinitely often. By Proposition 3.2.5 the set of points $\cup_n \{x_n\}$ has no limit point in K. (If it did, you would have a subsequence converging to this point since every ball containing this point would contain infinitely many points of $\cup_n \{x_n\}$.) Now consider the sets $H_n \equiv \cup_{k\geq n} \{x_k\} \cup H'$ where H' denotes all limit points of $\cup_n \{x_n\}$ in X which is the same as the limit points of $\cup_{k\geq n} \{x_k\}$. Therefore, each H_n is closed thanks to Lemma 3.3.2. Now let $U_n \equiv H_n^C$. This is an increasing sequence of open sets whose union contains K thanks to the fact that there is no constant subsequence. However, none of these open sets covers K because U_n is missing x_n, violating the definition of compactness. Next is an alternate argument.

$\Rightarrow$ Now suppose K is compact. If it is not sequentially compact, then there exists a sequence $\{x_n\}$ which has no convergent subsequence to a point of K. If $x \in K$, then there exists $B(x,r_x)$ which contains x_n for only finitely many n. This is because x is not the limit of a subsequence. Then $\{B(x_i,r_i)\}_{i=1}^N$ is a finite sub-cover of K. If p is the largest index for any x_k contained in $\cup_{i=1}^N B(x_i,r_i)$, let $n > p$ and consider x_n. It is a point in K but it can't be in any of the sets covering K. ■

Definition 3.5.6 *X be a metric space. Then a finite set of points $\{x_1,\cdots,x_n\}$ is called an ε net if $X\subseteq\cup_{k=1}^{n}B(x_k,\varepsilon)$. If, for every $\varepsilon>0$ a metric space has an ε net, then we say that the metric space is totally bounded.*

Lemma 3.5.7 *If a metric space (K,d) is sequentially compact, then it is separable and totally bounded.*

Proof: Pick $x_1\in K$. If $B(x_1,\varepsilon)\supseteq K$, then stop. Otherwise, pick $x_2\notin B(x_1,\varepsilon)$. Continue this way. If $\{x_1,\cdots,x_n\}$ have been chosen, either $K\subseteq\cup_{k=1}^{n}B(x_k,\varepsilon)$ in which case, you have found an ε net or this does not happen in which case, you can pick $x_{n+1}\notin\cup_{k=1}^{n}B(x_k,\varepsilon)$. The process must terminate since otherwise, the sequence would need to have a convergent subsequence which is not possible because every pair of terms is farther apart than ε. See Lemma 3.2.4. Thus for every $\varepsilon>0$, there is an ε net. Thus the metric space is totally bounded. Let N_ε denote an ε net. Let $D=\cup_{k=1}^{\infty}N_{1/2^k}$. Then this is a countable dense set. It is countable because it is the countable union of finite sets and it is dense because given a point, there is a point of D within $1/2^k$ of it. ■

Also recall that a complete metric space is one for which every Cauchy sequence converges to a point in the metric space.

The following is the main theorem which relates these concepts.

Theorem 3.5.8 *For (X,d) a metric space, the following are equivalent.*

1. *(X,d) is compact.*
2. *(X,d) is sequentially compact.*
3. *(X,d) is complete and totally bounded.*

Proof: By Theorem 3.5.5, the first two conditions are equivalent.

2.$\Longrightarrow$ 3. If (X,d) is sequentially compact, then by Lemma 3.5.7, it is totally bounded. If $\{x_n\}$ is a Cauchy sequence, then there is a subsequence which converges to $x\in X$ by assumption. However, from Theorem 3.2.2 this requires the original Cauchy sequence to converge.

3.$\Longrightarrow$ 1. Since (X,d) is totally bounded, there must be a countable dense subset of X. Just take the union of $1/2^k$ nets for each $k\in\mathbb{N}$. Thus (X,d) is completely separable by Theorem 3.4.6 has the Lindeloff property. Hence, if X is not compact, there is a countable set of open sets $\{U_i\}_{i=1}^{\infty}$ which covers X but no finite subset does. Consider the nonempty closed sets F_n and pick $x_n\in F_n$ where

$$X\setminus\cup_{i=1}^{n}U_i\equiv X\cap(\cup_{i=1}^{n}U_i)^C\equiv F_n$$

Let $\left\{x_m^k\right\}_{m=1}^{M_k}$ be a $1/2^k$ net for X. We have for some m, $B\left(x_{m_k}^k,1/2^k\right)$ contains x_n for infinitely many values of n because there are only finitely many balls and infinitely many indices. Then out of the finitely many $\left\{x_m^{k+1}\right\}$ where $B\left(x_m^{k+1},1/2^{k+1}\right)$ has nonempty intersection with $B\left(x_{m_k}^k,1/2^k\right)$, pick one $x_{m_{k+1}}^{k+1}$ such that $B\left(x_{m_{k+1}}^{k+1},1/2^{k+1}\right)$ contains x_n for infinitely many n. Then obviously $\left\{x_{m_k}^k\right\}_{k=1}^{\infty}$ is a Cauchy sequence because

$$d\left(x_{m_k}^k,x_{m_{k+1}}^{k+1}\right)\leq\frac{1}{2^k}+\frac{1}{2^{k+1}}\leq\frac{1}{2^{k-1}}$$

Hence for $p < q$,

$$d\left(x_{m_p}^p, x_{m_q}^q\right) \leq \sum_{k=p}^{q-1} d\left(x_{m_k}^k, x_{m_{k+1}}^{k+1}\right) < \sum_{k=p}^{\infty} \frac{1}{2^{k-1}} = \frac{1}{2^{p-2}}$$

Now take a subsequence $x_{n_k} \in B\left(x_{m_k}^k, 2^{-k}\right)$ so it follows that $\lim_{k\to\infty} x_{n_k} = \lim_{k\to\infty} x_{m_k}^k = x \in X$. However, $x \in F_n$ for each n since each F_n is closed and these sets are nested. Thus $x \in \cap_n F_n$ contrary to the claim that $\{U_i\}_{i=1}^{\infty}$ covers X. ■

For the sake of another point of view, here is another argument, this time that 3.)⇒ 2.). This will illustrate something called the Cantor diagonalization process.

Assume 3.). Suppose $\{x_k\}$ is a sequence in X. By assumption there are finitely many open balls of radius $1/n$ covering X. This for each $n \in \mathbb{N}$. Therefore, for $n = 1$, there is one of the balls, having radius 1 which contains x_k for infinitely many k. Therefore, there is a subsequence with every term contained in this ball of radius 1. Now do for this subsequence what was just done for $\{x_k\}$. There is a further subsequence contained in a ball of radius 1/2. Continue this way. Denote the i^{th} subsequence as $\{x_{ki}\}_{k=1}^{\infty}$. Arrange them as shown

$$\begin{array}{c} x_{11}, x_{21}, x_{31}, x_{41} \cdots \\ x_{12}, x_{22}, x_{32}, x_{42} \cdots \\ x_{13}, x_{23}, x_{33}, x_{43} \cdots \\ \vdots \end{array}$$

Thus all terms of $\{x_{ki}\}_{k=1}^{\infty}$ are contained in a ball of radius $1/i$. Consider now the diagonal sequence defined as $y_k \equiv x_{kk}$. Given n, each y_k is contained in a ball of radius $1/n$ whenever $k \geq n$. Thus $\{y_k\}$ is a subsequence of the original sequence and $\{y_k\}$ is a Cauchy sequence. By completeness of X, this converges to some $x \in X$ which shows that every sequence in X has a convergent subsequence. This shows 3.)⇒ 2.). ■

Lemma 3.5.9 *The closed interval $[a,b]$ in $\mathbb{R}$ is compact and every Cauchy sequence in $\mathbb{R}$ converges.*

Proof: To show this, suppose it is not. Then there is an open cover $\mathscr{C}$ which admits no finite subcover for $[a,b] \equiv I_0$. Consider the two intervals $\left[a, \frac{a+b}{2}\right], \left[\frac{a+b}{2}, b\right]$. One of these, maybe both cannot be covered with finitely many sets of $\mathscr{C}$ since otherwise, there would be a finite collection of sets from $\mathscr{C}$ covering $[a,b]$. Let I_1 be the interval which has no finite subcover. Now do for it what was done for I_0. Split it in half and pick the half which has no finite covering of sets of $\mathscr{C}$. Thus there is a "nested" sequence of closed intervals $I_0 \supseteq I_1 \supseteq I_2 \cdots$, each being half of the preceding interval. Say $I_n = [a_n, b_n]$. By the nested interval Lemma, Lemma 2.7.1, there is a point x in all these intervals. The point is unique because the lengths of the intervals converge to 0. This point is in some $O \in \mathscr{C}$. Thus for some $\delta > 0, [x-\delta, x+\delta]$, having length 2δ, is contained in O. For k large enough, the interval $[a_k, b_k]$ has length less than δ but contains x. Therefore, it is contained in $[x-\delta, x+\delta]$ and so must be contained in a single set of $\mathscr{C}$ contrary to the construction. This contradiction shows that in fact $[a,b]$ is compact.

Now if $\{x_n\}$ is a Cauchy sequence, then it is contained in some interval $[a,b]$ which is compact. Hence there is a subsequence which converges to some $x \in [a,b]$. By Theorem 3.2.2 the original Cauchy sequence converges to x. ■

3.6 Continuous Functions

The following is a fairly general definition of what it means for a function to be continuous. It includes everything seen in typical calculus classes as a special case.

Definition 3.6.1 *Let $f : X \to Y$ be a function where (X,d) and (Y,ρ) are metric spaces. Then f is continuous at $x \in X$ if and only if the following condition holds. For every $\varepsilon > 0$, there exists $\delta > 0$ such that if $d(\hat{x},x) < \delta$, then $\rho(f(\hat{x}), f(x)) < \varepsilon$. If f is continuous at every $x \in X$ we say that f is continuous on X.*

For example, you could have a real valued function $f(x)$ defined on an interval $[0,1]$. In this case you would have $X = [0,1]$ and $Y = \mathbb{R}$ with the distance given by $d(x,y) = |x-y|$. Then the following theorem is the main result.

Theorem 3.6.2 *Let $f : X \to Y$ where (X,d) and (Y,ρ) are metric spaces. Then the following two are equivalent.*

a f is continuous at x.

b Whenever $x_n \to x$, it follows that $f(x_n) \to f(x)$.

Also, the following are equivalent.

c f is continuous on X.

d Whenever V is open in Y, it follows that $f^{-1}(V) \equiv \{x : f(x) \in V\}$ is open in X.

e Whenever H is closed in Y, it follows that $f^{-1}(H) \equiv \{x : f(x) \in H\}$ is closed in X.

Proof: a $\Longrightarrow$ b: Let f be continuous at x and suppose $x_n \to x$. Then let $\varepsilon > 0$ be given. By continuity, there exists $\delta > 0$ such that if $d(\hat{x},x) < \delta$, then $\rho(f(\hat{x}), f(x)) < \varepsilon$. Since $x_n \to x$, it follows that there exists N such that if $n \geq N$, then $d(x_n,x) < \delta$ and so, if $n \geq N$, it follows that $\rho(f(x_n), f(x)) < \varepsilon$. Since $\varepsilon > 0$ is arbitrary, it follows that $f(x_n) \to f(x)$.

b $\Longrightarrow$ a: Suppose b holds but f fails to be continuous at x. Then there exists $\varepsilon > 0$ such that for all $\delta > 0$, there exists $\hat{x}$ such that $d(\hat{x},x) < \delta$ but $\rho(f(\hat{x}), f(x)) \geq \varepsilon$. Letting $\delta = 1/n$, there exists x_n such that $d(x_n,x) < 1/n$ but $\rho(f(x_n), f(x)) \geq \varepsilon$. Now this is a contradiction because by assumption, the fact that $x_n \to x$ implies that $f(x_n) \to f(x)$. In particular, for large enough n, $\rho(f(x_n), f(x)) < \varepsilon$ contrary to the construction.

c $\Longrightarrow$ d: Let V be open in Y. Let $x \in f^{-1}(V)$ so that $f(x) \in V$. Since V is open, there exists $\varepsilon > 0$ such that $B(f(x),\varepsilon) \subseteq V$. Since f is continuous at x, it follows that there exists $\delta > 0$ such that if $\hat{x} \in B(x,\delta)$, then $f(\hat{x}) \in B(f(x),\varepsilon) \subseteq V$. $(f(B(x,\delta)) \subseteq B(f(x),\varepsilon))$ In other words, $B(x,\delta) \subseteq f^{-1}(B(f(x),\varepsilon)) \subseteq f^{-1}(V)$ which shows that, since x was an arbitrary point of $f^{-1}(V)$, every point of $f^{-1}(V)$ is an interior point which implies $f^{-1}(V)$ is open.

d $\Longrightarrow$ e: Let H be closed in Y. Then $f^{-1}(H)^C = f^{-1}\left(H^C\right)$ which is open by assumption. Hence $f^{-1}(H)$ is closed because its complement is open.

e $\Longrightarrow$ d: Let V be open in Y. Then $f^{-1}(V)^C = f^{-1}\left(V^C\right)$ which is assumed to be closed. This is because the complement of an open set is a closed set.

d $\Longrightarrow$ c: Let $x \in X$ be arbitrary. Is it the case that f is continuous at x? Let $\varepsilon > 0$ be given. Then $B(f(x),\varepsilon)$ is an open set in V and so $x \in f^{-1}(B(f(x),\varepsilon))$ which is given to be open. Hence there exists $\delta > 0$ such that $x \in B(x,\delta) \subseteq f^{-1}(B(f(x),\varepsilon))$. Thus, $f(B(x,\delta)) \subseteq B(f(x),\varepsilon)$ so $\rho(f(\hat{x}), f(x)) < \varepsilon$. Thus f is continuous at x for every x. ■

Example 3.6.3 *$x \to d(x,y)$ is a continuous function from the metric space to the metric space of nonnegative real numbers.*

This follows from Lemma 3.2.6. You can also define a metric on a Cartesian product of metric spaces.

Proposition 3.6.4 *Let (X,d) be a metric space and consider $(X \times X, \rho)$ where*

$$\rho((x,\tilde{x}),(y,\tilde{y})) \equiv d(x,y) + d(\tilde{x},\tilde{y}).$$

Then this is also a metric space.

Proof: The only condition not obvious is the triangle inequality. However,

$$\rho((x,\tilde{x}),(y,\tilde{y})) + \rho((y,\tilde{y}),(z,\tilde{z})) \equiv d(x,y) + d(\tilde{x},\tilde{y}) + d(y,z) + d(\tilde{y},\tilde{z})$$

$$\geq d(x,z) + d(\tilde{x},\tilde{z}) = \rho((x,\tilde{x}),(z,\tilde{z})) \blacksquare$$

Definition 3.6.5 *If you have two metric spaces (X,d) and (Y,ρ), a function $f : X \to Y$ is called a homeomorphism if and only if it is continuous, one to one, onto, and its inverse is also continuous.*

Here is a useful proposition.

Proposition 3.6.6 *Let (X,d) be a metric space and let S be a nonempty subset of X. Define*

$$\operatorname{dist}(x,S) \equiv \inf\{d(x,s) : s \in S\}$$

Then $|\operatorname{dist}(x,S) - \operatorname{dist}(y,S)| \leq d(x,y)$ so $x \to \operatorname{dist}(x,S)$ is continuous.

Proof: Say $\operatorname{dist}(x,S) \geq \operatorname{dist}(y,S)$. Then there is $s \in S$ such that $\operatorname{dist}(y,S) + \varepsilon > d(y,s)$. Then

$$|\operatorname{dist}(x,S) - \operatorname{dist}(y,S)| = \operatorname{dist}(x,S) - \operatorname{dist}(y,S) \leq d(x,s) - (d(y,s) - \varepsilon)$$

$$\leq d(x,y) + d(y,s) - (d(y,s) - \varepsilon) = d(x,y) + \varepsilon$$

Since $\varepsilon > 0$ is arbitrary, this shows the claimed result. If $\operatorname{dist}(x,S) \leq \operatorname{dist}(y,S)$, repeat switching roles of x and y. $\blacksquare$

3.7 Continuity and Compactness

How does compactness relate to continuity? It turns out that the continuous image of a compact set is always compact. This is an easy consequence of the above major theorem.

Theorem 3.7.1 *Let $f : X \to Y$ where (X,d) and (Y,ρ) are metric spaces and f is continuous on X. Then if $K \subseteq X$ is compact, it follows that $f(K)$ is compact in (Y,ρ).*

Proof: Let $\mathscr{C}$ be an open cover of $f(K)$. Denote by $f^{-1}(\mathscr{C})$ the sets of the form $\{f^{-1}(U) : U \in \mathscr{C}\}$. Then $f^{-1}(\mathscr{C})$ is an open cover of K. It follows there are finitely many sets of the form $\{f^{-1}(U_1), \cdots, f^{-1}(U_n)\}$ which covers K. It follows that $\{U_1, \cdots, U_n\}$ is an open cover for $f(K)$. $\blacksquare$

The following is the important extreme values theorem for a real valued function defined on a compact set.

Theorem 3.7.2 *Let K be a compact metric space and suppose $f : K \to \mathbb{R}$ is a continuous function. That is, $\mathbb{R}$ is the metric space where the metric is given by $d(x,y) = |x-y|$. Then f achieves its maximum and minimum values on K.*

Proof: Let $\lambda \equiv \sup\{f(x) : x \in K\}$. Then from the definition of sup, you have the existence of a sequence $\{x_n\} \subseteq K$ such that $\lim_{n\to\infty} f(x_n) = \lambda$. There is a subsequence still called $\{x_n\}$ which converges to some $x \in K$. From continuity, $\lambda = \lim_{n\to\infty} f(x_n) = f(x)$ and so f achieves its maximum value at x. Similar reasoning shows that it achieves its minimum value on K. ■

Definition 3.7.3 *Let $f : (X,d) \to (Y,\rho)$ be a function. Then it is said to be uniformly continuous on X if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that whenever $x, \hat{x}$ are two points of X with $d(x,\hat{x}) < \delta$, it follows that $\rho(f(x), f(\hat{x})) < \varepsilon$.*

Note the difference between this and continuity. With continuity, the δ could depend on x but here it works for any pair of points in X.

There is a remarkable result concerning compactness and uniform continuity.

Theorem 3.7.4 *Let $f : (X,d) \to (Y,\rho)$ be a continuous function and let K be a compact subset of X. Then the restriction of f to K is uniformly continuous.*

Proof: First of all, K is a metric space and f restricted to K is continuous. Now suppose it fails to be uniformly continuous. Then there exists $\varepsilon > 0$ and pairs of points $x_n, \hat{x}_n$ such that $d(x_n, \hat{x}_n) < 1/n$ but $\rho(f(x_n), f(\hat{x}_n)) \geq \varepsilon$. Since K is compact, it is sequentially compact and so there exists a subsequence, still denoted as $\{x_n\}$ such that $x_n \to x \in K$. Then also $\hat{x}_n \to x$ also and so by Lemma 3.2.6, $\rho(f(x), f(x)) = \lim_{n\to\infty} \rho(f(x_n), f(\hat{x}_n)) \geq \varepsilon$ which is a contradiction. ■

3.8 Lipschitz Continuity and Contraction Maps

The following may be of more interest in the case of normed vector spaces, but there is no harm in stating it in this more general setting. You should verify that the functions described in the following definition are all continuous.

Definition 3.8.1 *Let $f : X \to Y$ where (X,d) and (Y,ρ) are metric spaces. Then f is said to be Lipschitz continuous if for every $x, \hat{x} \in X$, $\rho(f(x), f(\hat{x})) \leq r\, d(x,\hat{x})$. The function is called a contraction map if $r < 1$.*

The big theorem about contraction maps is the following.

Theorem 3.8.2 *Let $f : (X,d) \to (X,d)$ be a contraction map and let (X,d) be a complete metric space. Thus Cauchy sequences converge and also $d(f(x), f(\hat{x})) \leq r\, d(x,\hat{x})$ where $r < 1$. Then f has a unique fixed point. This is a point $x \in X$ such that $f(x) = x$. Also, if x_0 is any point of X, then*

$$d(x, x_0) \leq \frac{d(x_0, f(x_0))}{1-r}$$

Also, for each n,

$$d(f^n(x_0), x_0) \leq \frac{d(x_0, f(x_0))}{1-r},$$

and $x = \lim_{n\to\infty} f^n(x_0)$.

Proof: Pick $x_0 \in X$ and consider the sequence of the iterates of the map f given by $x_0, f(x_0), f^2(x_0), \cdots$. We argue that this is a Cauchy sequence. For $m < n$, it follows from the triangle inequality,

$$d(f^m(x_0), f^n(x_0)) \leq \sum_{k=m}^{n-1} d\left(f^{k+1}(x_0), f^k(x_0)\right) \leq \sum_{k=m}^{\infty} r^k d(f(x_0), x_0)$$

The reason for this last is as follows.

$$d\left(f^2(x_0), f(x_0)\right) \leq r d(f(x_0), x_0)$$

$$d\left(f^3(x_0), f^2(x_0)\right) \leq r d\left(f^2(x_0), f(x_0)\right) \leq r^2 d(f(x_0), x_0)$$

and so forth. Therefore, by the triangle inequality,

$$\begin{aligned} d(f^m(x_0), f^n(x_0)) &\leq \sum_{k=m}^{n-1} d\left(f^{k+1}(x_0), f^k(x_0)\right) \\ &\leq \sum_{k=m}^{\infty} r^k d(f(x_0), x_0) \leq d(f(x_0), x_0) \frac{r^m}{1-r} \end{aligned} \tag{3.1}$$

which shows that this is indeed a Cauchy sequence. Therefore, there exists x such that $\lim_{n\to\infty} f^n(x_0) = x$. By continuity, $f(x) = f(\lim_{n\to\infty} f^n(x_0)) = \lim_{n\to\infty} f^{n+1}(x_0) = x$.

Also note that, letting $m = 0$ in 3.1, this estimate yields

$$d(x_0, f^n(x_0)) \leq \frac{d(x_0, f(x_0))}{1-r}$$

Now $d(x_0, x) \leq d(x_0, f^n(x_0)) + d(f^n(x_0), x)$ and so

$$d(x_0, x) - d(f^n(x_0), x) \leq \frac{d(x_0, f(x_0))}{1-r}$$

Letting $n \to \infty$, it follows that $d(x_0, x) \leq \frac{d(x_0, f(x_0))}{1-r}$ because $\lim_{n\to\infty} d(f^n(x_0), x) = d(x, x) = 0$ by Lemma 3.2.6.

It only remains to verify that there is only one fixed point. Suppose then that x, x' are two. Then

$$d(x, x') = d\left(f(x), f\left(x'\right)\right) \leq r d\left(x', x\right)$$

and so $d(x, x') = 0$ because $r < 1$. ■

The above is the usual formulation of this important theorem, but we actually proved a better result.

Corollary 3.8.3 *Let B be a closed subset of the complete metric space (X, d) and let $f : B \to X$ be a contraction map*

$$d(f(x), f(\hat{x})) \leq r d(x, \hat{x}), \ r < 1.$$

Also suppose ***there exists*** *$x_0 \in B$ such that the sequence of iterates $\{f^n(x_0)\}_{n=1}^{\infty}$ remains in B. Then f has a unique fixed point in B which is the limit of the sequence of iterates. This is a point $x \in B$ such that $f(x) = x$. In the case that $B = \overline{B(x_0, \delta)}$, the sequence of iterates satisfies the inequality*

$$d(f^n(x_0), x_0) \leq \frac{d(x_0, f(x_0))}{1-r}$$

and so it will remain in B if $\frac{d(x_0, f(x_0))}{1-r} < \delta$.

Proof: By assumption, the sequence of iterates stays in B. Then, as in the proof of the preceding theorem, for $m<n$, it follows from the triangle inequality,

$$\begin{aligned} d\left(f^m(x_0), f^n(x_0)\right) &\leq \sum_{k=m}^{n-1} d\left(f^{k+1}(x_0), f^k(x_0)\right) \\ &\leq \sum_{k=m}^{\infty} r^k d\left(f(x_0), x_0\right) = \frac{r^m}{1-r} d\left(f(x_0), x_0\right) \end{aligned}$$

Hence the sequence of iterates is Cauchy and must converge to a point x in X. However, B is closed and so it must be the case that $x \in B$. Then as before,

$$x = \lim_{n\to\infty} f^n(x_0) = \lim_{n\to\infty} f^{n+1}(x_0) = f\left(\lim_{n\to\infty} f^n(x_0)\right) = f(x)$$

As to the sequence of iterates remaining in B where B is a ball as described, the inequality above in the case where $m=0$ yields $d\left(x_0, f^n(x_0)\right) \leq \frac{1}{1-r} d\left(f(x_0), x_0\right)$ and so, if the right side is less than δ, then the iterates remain in B. As to the fixed point being unique, it is as before. If x, x' are both fixed points in B, then $d(x,x') = d\left(f(x), f(x')\right) \leq r d(x,x')$ and so $x = x'$. ■

The contraction mapping theorem has an extremely useful generalization. In order to get a unique fixed point, it suffices to have some power of f a contraction map.

Theorem 3.8.4 *Let $f : (X,d) \to (X,d)$ have the property that for some $n \in \mathbb{N}$, f^n is a contraction map and let (X,d) be a complete metric space. Then there is a unique fixed point for f. As in the earlier theorem the sequence of iterates $\{f^n(x_0)\}_{n=1}^{\infty}$ also converges to the fixed point.*

Proof: From Theorem 3.8.2 there is a unique fixed point for f^n. Thus $f^n(x) = x$ Then

$$f^n\left(f(x)\right) = f^{n+1}(x) = f(x)$$

By uniqueness, $f(x) = x$.

Now consider the sequence of iterates. Suppose it fails to converge to x. Then there is $\varepsilon > 0$ and a subsequence n_k such that $d\left(f^{n_k}(x_0), x\right) \geq \varepsilon$. Now $n_k = p_k n + r_k$ where r_k is one of the numbers $\{0, 1, 2, \cdots, n-1\}$. It follows that there exists one of these numbers which is repeated infinitely often. Call it r and let the further subsequence continue to be denoted as n_k. Thus $d\left(f^{p_k n + r}(x_0), x\right) \geq \varepsilon$. In other words,

$$d\left(f^{p_k n}\left(f^r(x_0)\right), x\right) \geq \varepsilon$$

However, from Theorem 3.8.2, as $k \to \infty$, $f^{p_k n}\left(f^r(x_0)\right) \to x$ which contradicts the above inequality. Hence the sequence of iterates converges to x, as it did for f a contraction map. ■

3.9 Convergence of Functions

Next is to consider the meaning of convergence of sequences of functions. There are two main ways of convergence of interest here, pointwise and uniform convergence.

Definition 3.9.1 *Let $f_n : X \to Y$ where $(X,d), (Y,\rho)$ are two metric spaces. Then $\{f_n\}$ is said to converge pointwise to a function $f : X \to Y$ if for every $x \in X$, $\lim_{n\to\infty} f_n(x) = f(x)$. $\{f_n\}$ is said to converge uniformly if for all $\varepsilon > 0$, there exists N such that if $n \geq N$, then $\sup_{x\in X} \rho\left(f_n(x), f(x)\right) < \varepsilon$*

Here is a well known example illustrating the difference between pointwise and uniform convergence.

Example 3.9.2 *Let $f_n(x) = x^n$ on the metric space $[0,1]$. Then this function converges pointwise to*

$$f(x) = \begin{cases} 0 \text{ on } [0,1) \\ 1 \text{ at } 1 \end{cases}$$

but it does not converge uniformly on this interval to f.

Note how the target function f in the above example is not continuous even though each function in the sequence is. The nice thing about uniform convergence is that it takes continuity of the functions in the sequence and imparts it to the target function. It does this for both continuity at a single point and uniform continuity. Thus uniform convergence is a very superior thing.

Theorem 3.9.3 *Let $f_n : X \to Y$ where $(X,d),(Y,\rho)$ are two metric spaces and suppose each f_n is continuous at $x \in X$ and also that f_n converges uniformly to f on X. Then f is also continuous at x. In addition to this, if each f_n is uniformly continuous on X, then the same is true for f.*

Proof: Let $\varepsilon > 0$ be given. Then

$$\rho(f(x),f(\hat{x})) \leq \rho(f(x),f_n(x)) + \rho(f_n(x),f_n(\hat{x})) + \rho(f_n(\hat{x}),f(\hat{x}))$$

By uniform convergence, there exists N such that both $\rho(f(x),f_n(x)), \rho(f_n(\hat{x}),f(\hat{x}))$ are less than $\varepsilon/3$ provided $n \geq N$. Thus picking such an n

$$\rho(f(x),f(\hat{x})) \leq \frac{2\varepsilon}{3} + \rho(f_n(x),f_n(\hat{x}))$$

From the continuity of f_n, there exists a positive number $\delta > 0$ such that if $d(x,\hat{x}) < \delta$, then $\rho(f_n(x),f_n(\hat{x})) < \varepsilon/3$. Hence, if $d(x,\hat{x}) < \delta$, then

$$\rho(f(x),f(\hat{x})) \leq \frac{2\varepsilon}{3} + \rho(f_n(x),f_n(\hat{x})) < \frac{2\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon$$

Hence, f is continuous at x.

Next consider uniform continuity. It follows from the uniform convergence that if $x,\hat{x}$ are any two points of X, then if $n \geq N$, then, picking such an $n, \rho(f(x),f(\hat{x})) \leq \frac{2\varepsilon}{3} + \rho(f_n(x),f_n(\hat{x}))$. By uniform continuity of f_n there exists δ such that if $d(x,\hat{x}) < \delta$, then the term on the right in the above is less than $\varepsilon/3$. Hence if $d(x,\hat{x}) < \delta$, then $\rho(f(x),f(\hat{x})) < \varepsilon$ and so f is uniformly continuous as claimed. ■

3.10 Compactness in $C(X,Y)$ Ascoli Arzela Theorem

This will use the characterization of compact metric spaces to give a proof of a general version of the Arzella Ascoli theorem. See Naylor and Sell [42] which is where I saw this general formulation.

Definition 3.10.1 *Let (X,d_X) be a compact metric space. Let (Y,d_Y) be another complete metric space. Then $C(X,Y)$ will denote the continuous functions which map X to Y. Then ρ is a metric on $C(X,Y)$ defined by $\rho(f,g) \equiv \sup_{x\in X} d_Y(f(x),g(x))$.*

Theorem 3.10.2 *$(C(X,Y),\rho)$ is a complete metric space where (X,d_X) is a compact metric space*

Proof: It is first necessary to show that ρ is well defined. In this argument, I will just write d rather than d_X or d_Y. To show this, note that from Lemma 3.2.6, if $x_n \to x$, and $y_n \to y$, then $d(x_n,y_n) \to d(x,y)$. Therefore, if f,g are continuous, and $x_n \to x$ so $f(x_n) \to f(x)$ and $g(x_n) \to g(x)$, $d(f(x_n),g(x_n)) \to d(f(x),g(x))$ and so, $\rho(f,g)$ is just the maximum of a continuous function defined on a compact set. By Theorem 3.7.2, the extreme values theorem, this maximum exists.

Clearly $\rho(f,g) = \rho(g,f)$ and

$$\begin{aligned}\rho(f,g)+\rho(g,h) &= \sup_{x\in X} d(f(x),g(x)) + \sup_{x\in X} d(g(x),h(x)) \\ &\geq \sup_{x\in X} (d(f(x),g(x)) + d(g(x),h(x))) \\ &\geq \sup_{x\in X} (d(f(x),h(x))) = \rho(f,h)\end{aligned}$$

so the triangle inequality holds.

It remains to check completeness. Let $\{f_n\}$ be a Cauchy sequence. Then from the definition, $\{f_n(x)\}$ is a Cauchy sequence in Y and so it converges to something called $f(x)$. By Theorem 3.9.3, f is continuous. It remains to show that $\rho(f_n,f) \to 0$. Let $x \in X$. Then from what was just noted,

$$d(f_n(x),f(x)) = \lim_{m\to\infty} d(f_n(x),f_m(x)) \leq \limsup_{m\to\infty} \rho(f_n,f_m)$$

since $\{f_n\}$ is given to be a Cauchy sequence, there exists N such that if $m,n > N$, then $\rho(f_n,f_m) < \varepsilon$. Therefore, if $n > N, d(f_n(x),f(x)) \leq \limsup_{m\to\infty} \rho(f_n,f_m) \leq \varepsilon$. Since x is arbitrary, it follows that $\rho(f_n,f) \leq \varepsilon$, if $n \geq N$. ■

Here is a useful lemma.

Lemma 3.10.3 *Let S be a totally bounded subset of (X,d) a metric space. Then $\overline{S}$ is also totally bounded.*

Proof: Suppose not. Then there exists a sequence $\{p_n\} \subseteq \overline{S}$ such that

$$d(p_m,p_n) \geq \varepsilon$$

for all $m \neq n$. Now let $q_n \in B\left(p_n,\frac{\varepsilon}{8}\right) \cap S$. Then it follows that

$$\frac{\varepsilon}{8} + d(q_n,q_m) + \frac{\varepsilon}{8} \geq d(p_n,q_n) + d(q_n,q_m) + d(q_m,p_m) \geq d(p_n,q_m) \geq \varepsilon$$

and so $d(q_n,q_m) > \frac{\varepsilon}{2}$. This contradicts total boundedness of S. ■

Next, here is an important definition.

Definition 3.10.4 *Let $\mathscr{A} \subseteq C(X,Y)$ where (X,d_X) and (Y,d_Y) are metric spaces. Thus $\mathscr{A}$ is a set of continuous functions mapping X to Y. Then $\mathscr{A}$ is said to be equicontinuous if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that if $d_X(x_1,x_2) < \delta$ then for all $f \in \mathscr{A}$, $d_Y(f(x_1),f(x_2)) < \varepsilon$. (This is uniform continuity which is uniform in $\mathscr{A}$.) $\mathscr{A}$ is said to be pointwise compact if $\{f(x) : f \in \mathscr{A}\}$ has compact closure in Y.*

Here is the Ascoli Arzela theorem.

Theorem 3.10.5 *Let (X,d_X) be a compact metric space and let (Y,d_Y) be a complete metric space. Thus $(C(X,Y),\rho)$ is a complete metric space. Let $\mathscr{A} \subseteq C(X,Y)$ be pointwise compact and equicontinuous. Then $\overline{\mathscr{A}}$ is compact. Here the closure is taken in $(C(X,Y),\rho)$.*

Proof: The more useful direction is that the two conditions imply compactness of $\overline{\mathscr{A}}$. I prove this first. Since $\overline{\mathscr{A}}$ is a closed subset of a complete space, it follows from Theorem 3.5.8, that $\overline{\mathscr{A}}$ will be compact if it is totally bounded. In showing this, it follows from Lemma 3.10.3 that it suffices to verify that $\mathscr{A}$ is totally bounded. Suppose this is not so. Then there exists $\varepsilon > 0$ and a sequence of points of $\mathscr{A}$, $\{f_n\}$ such that $\rho(f_n,f_m) \geq \varepsilon$ whenever $n \neq m$.

By equicontinuity, there exists $\delta > 0$ such that if

$$d(x,y) < \delta,$$

then $d_Y(f(x),f(y)) < \frac{\varepsilon}{8}$ for all $f \in \mathscr{A}$. Let $\{x_i\}_{i=1}^p$ be a δ net for X. Since there are only finitely many x_i, it follows from pointwise compactness that there exists a subsequence, still denoted by $\{f_n\}$ which converges at each x_i. Now let $x \in X$ be arbitrary. There exists N such that for each x_i in that δ net,

$$d_Y(f_n(x_i),f_m(x_i)) < \varepsilon/8 \text{ whenever } n,m \geq N$$

Then for $m,n \geq N$,

$$\begin{aligned} & d_Y(f_n(x),d_{Ym}(x)) \\ \leq\ & d_Y(f_n(x),f_n(x_i)) + d_Y(f_n(x_i),f_m(x_i)) + d_Y(f_m(x_i),f_m(x)) \\ <\ & d_Y(f_n(x),f_n(x_i)) + \varepsilon/8 + d_Y(f_m(x_i),f_m(x)) \end{aligned}$$

Pick x_i such that $d(x,x_i) < \delta$. $\{x_i\}_{i=1}^p$ is a δ net and so this is surely possible. Then by equicontinuity, the two ends are each less than $\varepsilon/8$ and so for $m,n \geq N$,

$$d_Y(f_n(x),f_m(x)) \leq \frac{3\varepsilon}{8}$$

Since x is arbitrary, it follows that $\rho(f_n,f_m) \leq 3\varepsilon/8 < \varepsilon$ which is a contradiction. It follows that $\mathscr{A}$ and hence $\overline{\mathscr{A}}$ is totally bounded. This proves the more important direction.

Next suppose $\overline{\mathscr{A}}$ is compact. Why must $\mathscr{A}$ be pointwise compact and equicontinuous? If it fails to be pointwise compact, then there exists $x \in X$ such that $\{f(x) : f \in \mathscr{A}\}$ is not contained in a compact set of Y. Thus there exists $\varepsilon > 0$ and a sequence of functions in $\mathscr{A}$ $\{f_n\}$ such that $d(f_n(x),f_m(x)) \geq \varepsilon$. But this implies $\rho(f_m,f_n) \geq \varepsilon$ and so $\overline{\mathscr{A}}$ fails to be totally bounded, a contradiction. Thus $\mathscr{A}$ must be pointwise compact. Now why must it be equicontinuous? If it is not, then for each $n \in \mathbb{N}$ there exists $\varepsilon > 0$ and $x_n,y_n \in X$ such that $d(x_n,y_n) < 1/n$ but for some $f_n \in \mathscr{A}$, $d(f_n(x_n),f_n(y_n)) \geq \varepsilon$. However, by compactness, there exists a subsequence $\{f_{n_k}\}$ such that $\lim_{k\to\infty} \rho(f_{n_k},f) = 0$ and also that $x_{n_k},y_{n_k} \to x \in X$. Hence

$$\begin{aligned} \varepsilon \leq\ & d(f_{n_k}(x_{n_k}),f_{n_k}(y_{n_k})) \leq d(f_{n_k}(x_{n_k}),f(x_{n_k})) \\ & + d(f(x_{n_k}),f(y_{n_k})) + d(f(y_{n_k}),f_{n_k}(y_{n_k})) \\ & \leq \rho(f_{n_k},f) + d(f(x_{n_k}),f(y_{n_k})) + \rho(f,f_{n_k}) \end{aligned}$$

and now this is a contradiction because each term on the right converges to 0. The middle term converges to 0 because $f(x_{n_k}),f(y_{n_k}) \to f(x)$. See Lemma 3.2.6. ■

3.11 Connected Sets

Stated informally, connected sets are those which are in one piece. In order to define what is meant by this, I will first consider what it means for a set to **not** be in one piece. This is called **separated.** Connected sets are defined in terms of **not** being separated. This is why theorems about connected sets sometimes seem a little tricky.

Definition 3.11.1 *A set, S in a metric space, is separated if there exist sets A, B such that*

$$S = A \cup B,\ A, B \neq \emptyset,\ \text{and } \overline{A} \cap B = \overline{B} \cap A = \emptyset.$$

In this case, the sets A and B are said to separate S. A set is connected if it is not separated. Remember $\overline{A}$ denotes the closure of the set A.

Note that the concept of connected sets is defined in terms of what it is not. This makes it somewhat difficult to understand. One of the most important theorems about connected sets is the following.

Theorem 3.11.2 *Suppose $\mathscr{U}$ is a set of connected sets and that there exists a point p which is in all of these connected sets. Then $K \equiv \cup \mathscr{U}$ is connected.*

Proof: The argument is dependent on Lemma 3.3.2. Suppose

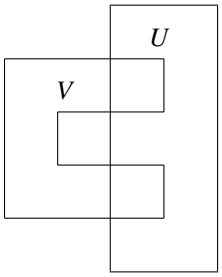
$$K = A \cup B$$

where $\bar{A} \cap B = \bar{B} \cap A = \emptyset, A \neq \emptyset, B \neq \emptyset$. Then p is in one of these sets. Say $p \in A$. Then if $U \in \mathscr{U}$, it must be the case that $U \subseteq A$ since if not, you would have

$$U = (A \cap U) \cup (B \cap U)$$

and the limit points of $A \cap U$ cannot be in B hence not in $B \cap U$ while the limit points of $B \cap U$ cannot be in A hence not in $A \cap U$. Thus $B = \emptyset$. It follows that K cannot be separated and so it is connected. ■

The intersection of connected sets is not necessarily connected as is shown by the following picture.



Theorem 3.11.3 *Let $f : X \to Y$ be continuous where Y is a metric space and X is connected. Then $f(X)$ is also connected.*

Proof: To do this you show $f(X)$ is not separated. Suppose to the contrary that $f(X) = A \cup B$ where A and B separate $f(X)$. Then consider the sets $f^{-1}(A)$ and $f^{-1}(B)$. If $z \in f^{-1}(B)$, then $f(z) \in B$ and so $f(z)$ is not a limit point of A. Therefore, there exists an

open set, U containing $\boldsymbol{f}(\boldsymbol{z})$ such that $U \cap A = \emptyset$. But then, the continuity of $\boldsymbol{f}$ and Theorem 3.6.2 implies that $\boldsymbol{f}^{-1}(U)$ is an open set containing $\boldsymbol{z}$ such that $\boldsymbol{f}^{-1}(U) \cap \boldsymbol{f}^{-1}(A) = \emptyset$. Therefore, $\boldsymbol{f}^{-1}(B)$ contains no limit points of $\boldsymbol{f}^{-1}(A)$. Similar reasoning implies $\boldsymbol{f}^{-1}(A)$ contains no limit points of $\boldsymbol{f}^{-1}(B)$. It follows that X is separated by $\boldsymbol{f}^{-1}(A)$ and $\boldsymbol{f}^{-1}(B)$, contradicting the assumption that X was connected. ■

An arbitrary set can be written as a union of maximal connected sets called connected components. This is the concept of the next definition.

Definition 3.11.4 *Let S be a set and let $\boldsymbol{p} \in S$. Denote by $C_{\boldsymbol{p}}$ the union of all connected subsets of S which contain $\boldsymbol{p}$. This is called the connected component determined by $\boldsymbol{p}$.*

Theorem 3.11.5 *Let $C_{\boldsymbol{p}}$ be a connected component of a set S in a metric space. Then $C_{\boldsymbol{p}}$ is a connected set and if $C_{\boldsymbol{p}} \cap C_{\boldsymbol{q}} \neq \emptyset$, then $C_{\boldsymbol{p}} = C_{\boldsymbol{q}}$.*

Proof: Let $\mathscr{C}$ denote the connected subsets of S which contain $\boldsymbol{p}$. By Theorem 3.11.2, $\cup\mathscr{C} = C_{\boldsymbol{p}}$ is connected. If $\boldsymbol{x} \in C_{\boldsymbol{p}} \cap C_{\boldsymbol{q}}$, then from Theorem 3.11.2, $C_{\boldsymbol{p}} \supseteq C_{\boldsymbol{p}} \cup C_{\boldsymbol{q}}$ and so $C_{\boldsymbol{p}} \supseteq C_{\boldsymbol{q}}$. The inclusion goes the other way by the same reason. ■

This shows the connected components of a set are equivalence classes and partition the set.

A set, I is an interval in $\mathbb{R}$ if and only if whenever $x, y \in I$ then $[x,y] \subseteq I$. The following theorem is about the connected sets in $\mathbb{R}$.

Theorem 3.11.6 *A set C in $\mathbb{R}$ is connected if and only if C is an interval.*

Proof: Let C be connected. If C consists of a single point, p, there is nothing to prove. The interval is just $[p,p]$. Suppose $p < q$ and $p, q \in C$. You need to show $(p,q) \subseteq C$. If

$$x \in (p,q) \setminus C$$

let $C \cap (-\infty, x) \equiv A$, and $C \cap (x, \infty) \equiv B$. Then $C = A \cup B$ and the sets A and B separate C contrary to the assumption that C is connected.

Conversely, let I be an interval. Suppose I is separated by A and B. Pick $x \in A$ and $y \in B$. Suppose without loss of generality that $x < y$. Now define the set,

$$S \equiv \{t \in [x,y] : [x,t] \subseteq A\}$$

and let l be the least upper bound of S. Then $l \in \overline{A}$ so $l \notin B$ which implies $l \in A$. But if $l \notin \overline{B}$, then for some $\delta > 0$,

$$(l, l+\delta) \cap B = \emptyset$$

contradicting the definition of l as an upper bound for S. Therefore, $l \in \overline{B}$ which implies $l \notin A$ after all, a contradiction. It follows I must be connected. ■

This yields a generalization of the intermediate value theorem from one variable calculus.

Corollary 3.11.7 *Let E be a connected set in a metric space and suppose $f : E \to \mathbb{R}$ and that $y \in (f(e_1), f(e_2))$ where $e_i \in E$. Then there exists $e \in E$ such that $f(e) = y$.*

Proof: From Theorem 3.11.3, $f(E)$ is a connected subset of $\mathbb{R}$. By Theorem 3.11.6 $f(E)$ must be an interval. In particular, it must contain y. This proves the corollary. ■

The following theorem is a very useful description of the open sets in $\mathbb{R}$.

Theorem 3.11.8 *Let U be an open set in $\mathbb{R}$. Then there exist countably many disjoint open sets $\{(a_i,b_i)\}_{i=1}^{\infty}$ such that $U=\cup_{i=1}^{\infty}(a_i,b_i)$.*

Proof: Let $p\in U$ and let $z\in C_p$, the connected component determined by p. Since U is open, there exists, $\delta>0$ such that $(z-\delta,z+\delta)\subseteq U$. It follows from Theorem 3.11.2 that

$$(z-\delta,z+\delta)\subseteq C_p.$$

This shows C_p is open. By Theorem 3.11.6, this shows C_p is an open interval, (a,b) where $a,b\in[-\infty,\infty]$. There are therefore at most countably many of these connected components because each must contain a rational number and the rational numbers are countable. Denote by $\{(a_i,b_i)\}_{i=1}^{\infty}$ the set of these connected components. ■

Definition 3.11.9 *A set E in a metric space is arcwise connected if for any two points, $\boldsymbol{p},\boldsymbol{q}\in E$, there exists a closed interval, $[a,b]$ and a continuous function, $\boldsymbol{\gamma}:[a,b]\to E$ such that $\boldsymbol{\gamma}(a)=p$ and $\boldsymbol{\gamma}(b)=q$.*

An example of an arcwise connected metric space would be any subset of $\mathbb{R}^n$ which is the continuous image of an interval. Arcwise connected is not the same as connected. A well known example is the following.

$$\left\{\left(x,\sin\frac{1}{x}\right):x\in(0,1]\right\}\cup\{(0,y):y\in[-1,1]\} \tag{3.2}$$

You can verify that this set of points in the normed vector space $\mathbb{R}^2$ is not arcwise connected but is connected.

Lemma 3.11.10 *In $\mathbb{R}^p$, $B(\boldsymbol{z},r)$ is arcwise connected.*

Proof: This is easy from the convexity of the set. If $\boldsymbol{x},\boldsymbol{y}\in B(\boldsymbol{z},r)$, then let $\boldsymbol{\gamma}(t)=\boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x})$ for $t\in[0,1]$.

$$\begin{aligned}\|\boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x})-\boldsymbol{z}\| &= \|(1-t)(\boldsymbol{x}-\boldsymbol{z})+t(\boldsymbol{y}-\boldsymbol{z})\|\\ &\leq (1-t)\|\boldsymbol{x}-\boldsymbol{z}\|+t\|\boldsymbol{y}-\boldsymbol{z}\|\\ &< (1-t)r+tr=r\end{aligned}$$

showing $\boldsymbol{\gamma}(t)$ stays in $B(\boldsymbol{z},r)$.■

Proposition 3.11.11 *If $X\neq\emptyset$ is arcwise connected, then it is connected.*

Proof: Let $p\in X$. Then by assumption, for any $x\in X$, there is an arc joining p and x. This arc is connected because it is the continuous image of an interval which is connected. Since x is arbitrary, every x is in a connected subset of X which contains p. Hence $C_p=X$ and so X is connected. ■

Theorem 3.11.12 *Let U be an open subset of $\mathbb{R}^p$. Then U is arcwise connected if and only if U is connected. Also the connected components of an open set are open sets.*

Proof: By Proposition 3.11.11 it is only necessary to verify that if U is connected and open, then U is arcwise connected. Pick $\boldsymbol{p} \in U$. Say $\boldsymbol{x} \in U$ satisfies $\mathscr{P}$ if there exists a continuous function, $\boldsymbol{\gamma} : [a,b] \to U$ such that $\boldsymbol{\gamma}(a) = \boldsymbol{p}$ and $\boldsymbol{\gamma}(b) = \boldsymbol{x}$.

$$A \equiv \{\boldsymbol{x} \in U \text{ such that } \boldsymbol{x} \text{ satisfies } \mathscr{P}.\}$$

If $\boldsymbol{x} \in A$, then Lemma 3.11.10 implies $B(\boldsymbol{x},r) \subseteq U$ is arcwise connected for small enough r. Thus letting $\boldsymbol{y} \in B(\boldsymbol{x},r)$, there exist intervals, $[a,b]$ and $[c,d]$ and continuous functions having values in U, $\boldsymbol{\gamma},\boldsymbol{\eta}$ such that $\boldsymbol{\gamma}(a) = \boldsymbol{p}, \boldsymbol{\gamma}(b) = \boldsymbol{x}, \boldsymbol{\eta}(c) = \boldsymbol{x}$, and $\boldsymbol{\eta}(d) = \boldsymbol{y}$. Then let $\boldsymbol{\gamma}_1 : [a, b+d-c] \to U$ be defined as

$$\boldsymbol{\gamma}_1(t) \equiv \begin{cases} \boldsymbol{\gamma}(t) \text{ if } t \in [a,b] \\ \boldsymbol{\eta}(t+c-b) \text{ if } t \in [b, b+d-c] \end{cases}$$

Then it is clear that $\boldsymbol{\gamma}_1$ is a continuous function mapping $\boldsymbol{p}$ to $\boldsymbol{y}$ and showing that $B(\boldsymbol{x},r) \subseteq A$. Therefore, A is open. $A \neq \emptyset$ because since U is open there is an open set, $B(\boldsymbol{p},\delta)$ containing $\boldsymbol{p}$ which is contained in U and is arcwise connected.

Now consider $B \equiv U \setminus A$. I claim this is also open. If B is not open, there exists a point $\boldsymbol{z} \in B$ such that every open set containing $\boldsymbol{z}$ is not contained in B. Therefore, letting $B(\boldsymbol{z},\delta)$ be such that $\boldsymbol{z} \in B(\boldsymbol{z},\delta) \subseteq U$, there exist points of A contained in $B(\boldsymbol{z},\delta)$. But then, a repeat of the above argument shows $z \in A$ also. Hence B is open and so if $B \neq \emptyset$, then $U = B \cup A$ and so U is separated by the two sets B and A contradicting the assumption that U is connected. Note that, since B is open, it contains no limit points of A and since A is open, it contains no limit points of B.

It remains to verify the connected components are open. Let $\boldsymbol{z} \in C_{\boldsymbol{p}}$ where $C_{\boldsymbol{p}}$ is the connected component determined by $\boldsymbol{p}$. Then picking $B(\boldsymbol{z},\delta) \subseteq U$, $C_{\boldsymbol{p}} \cup B(\boldsymbol{z},\delta)$ is connected and contained in U and so it must also be contained in $C_{\boldsymbol{p}}$. Thus $\boldsymbol{z}$ is an interior point of $C_{\boldsymbol{p}}$. ■

As an application, consider the following corollary.

Corollary 3.11.13 *Let $f : \Omega \to \mathbb{Z}$ be continuous where Ω is a connected nonempty open set of a metric space. Then f must be a constant.*

Proof: Suppose not. Then it achieves two different values, k and $l \neq k$. Then $\Omega = f^{-1}(l) \cup f^{-1}(\{m \in \mathbb{Z} : m \neq l\})$ and these are disjoint nonempty open sets which separate Ω. To see they are open, note

$$f^{-1}(\{m \in \mathbb{Z} : m \neq l\}) = f^{-1}\left(\cup_{m \neq l}\left(m - \frac{1}{6}, m + \frac{1}{6}\right)\right)$$

which is the inverse image of an open set while $f^{-1}(l) = f^{-1}\left(\left(l - \frac{1}{6}, l + \frac{1}{6}\right)\right)$ also an open set. ■

3.12 Partitions of Unity in Metric Space

Lemma 3.12.1 *Let X be a metric space and let S be a nonempty subset of X.*

$$\text{dist}(\boldsymbol{x},S) \equiv \inf\{d(\boldsymbol{x},\boldsymbol{z}) : \boldsymbol{z} \in S\}$$

Then

$$|\text{dist}(\boldsymbol{x},S) - \text{dist}(\boldsymbol{y},S)| \leq d(\boldsymbol{x},\boldsymbol{y}).$$

Proof: Say $\text{dist}(\boldsymbol{x},S) \geq \text{dist}(\boldsymbol{y},S)$. Then letting $\boldsymbol{\varepsilon} > 0$ be given, there exists $\boldsymbol{z} \in S$ such that $d(\boldsymbol{y},\boldsymbol{z}) < \text{dist}(\boldsymbol{y},S) + \boldsymbol{\varepsilon}$ Then

$$|\text{dist}(\boldsymbol{x},S) - \text{dist}(\boldsymbol{y},S)| = \text{dist}(\boldsymbol{x},S) - \text{dist}(\boldsymbol{y},S) \leq \text{dist}(\boldsymbol{x},S) - (d(\boldsymbol{y},\boldsymbol{z}) - \boldsymbol{\varepsilon})$$

$$\leq d(\boldsymbol{x},\boldsymbol{z}) - (d(\boldsymbol{y},\boldsymbol{z}) - \boldsymbol{\varepsilon}) \leq d(\boldsymbol{x},\boldsymbol{y}) + d(\boldsymbol{y},\boldsymbol{z}) - d(\boldsymbol{y},\boldsymbol{z}) + \boldsymbol{\varepsilon} = d(\boldsymbol{x},\boldsymbol{y}) + \boldsymbol{\varepsilon}$$

Since $\boldsymbol{\varepsilon}$ is arbitrary, $|\text{dist}(\boldsymbol{x},S) - \text{dist}(\boldsymbol{y},S)| \leq d(\boldsymbol{x},\boldsymbol{y})$. The situation is completely similar if $\text{dist}(\boldsymbol{x},S) < \text{dist}(\boldsymbol{y},S)$. ■

Then this shows that $\boldsymbol{x} \to \text{dist}(\boldsymbol{x},S)$ is a continuous real valued function.

This is about partitions of unity in metric space. Assume here that closed balls are compact. For example, you might be considering $\mathbb{R}^p$ with $d(\boldsymbol{x},\boldsymbol{y}) \equiv |\boldsymbol{x}-\boldsymbol{y}|$.

Definition 3.12.2 *Define* $\text{spt}(f)$ *(support of* f*) to be the closure of the set* $\{x : f(x) \neq 0\}$. *If* V *is an open set,* $C_c(V)$ *will be the set of continuous functions* f*, defined on* Ω *having* $\text{spt}(f) \subseteq V$.

Definition 3.12.3 *If* K *is a compact subset of an open set,* V*, then* $K \prec \phi \prec V$ *if*

$$\phi \in C_c(V),\ \phi(K) = \{1\},\ \phi(\Omega) \subseteq [0,1],$$

where Ω *denotes the whole metric space. Also for* $\phi \in C_c(\Omega)$*,* $K \prec \phi$ *if*

$$\phi(\Omega) \subseteq [0,1] \text{ and } \phi(K) = 1.$$

and $\phi \prec V$ *if*

$$\phi(\Omega) \subseteq [0,1] \text{ and } \text{spt}(\phi) \subseteq V.$$

Lemma 3.12.4 *Let* (Ω,d) *be a metric space in which closed balls are compact. Then if* K *is a compact subset of an open set* V*, then there exists* ϕ *such that* $K \prec \phi \prec V$.

Proof: Since K is compact, the distance between K and V^C is positive, $\delta > 0$. Otherwise there would be $x_n \in K$ and $y_n \in V^C$ with $d(x_n,y_n) < 1/n$. Taking a subsequence, still denoted with n, we can assume $x_n \to x$ and $y_n \to x$ but this would imply x is in both K and V^C which is not possible. Now consider $\{B(x,\delta/2)\}$ for $x \in K$. This is an open cover and the closure of each ball is contained in V. Since K is compact, finitely many of these balls cover K. Denote their union as W. Then $\overline{W}$ is compact because it is the finite union of the closed balls. Hence $K \subseteq W \subseteq \overline{W} \subseteq V$. Now consider

$$\phi(x) \equiv \frac{\text{dist}\left(x,W^C\right)}{\text{dist}(x,K) + \text{dist}(x,W^C)}$$

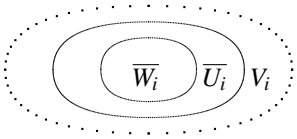
the denominator is never zero because x cannot be in both K and W^C. Thus ϕ is continuous by Lemma 3.12.1. also if $x \in K$, then $\phi(x) = 1$ and if $x \notin W$, then $\phi(x) = 0$. ■

Theorem 3.12.5 *(Partition of unity) Let* K *be a compact subset of a metric space in which closed balls are compact and suppose* $K \subseteq V = \cup_{i=1}^n V_i$*,* V_i *open. Then there exist* $\psi_i \prec V_i$ *with* $\sum_{i=1}^n \psi_i(x) = 1$ *for all* $x \in K$.

Proof: Let $K_1 = K \setminus \cup_{i=2}^n V_i$. Thus K_1 is compact and $K_1 \subseteq V_1$. Let $K_1 \subseteq W_1 \subseteq \overline{W}_1 \subseteq V_1$ with $\overline{W}_1$ compact. To obtain W_1, use Lemma 3.12.4 to get f such that $K_1 \prec f \prec V_1$ and let $W_1 \equiv \{x : f(x) \neq 0\}$. Thus $W_1, V_2, \cdots V_n$ covers K and $\overline{W}_1 \subseteq V_1$. Let $K_2 = K \setminus (\cup_{i=3}^n V_i \cup W_1)$. Then K_2 is compact and $K_2 \subseteq V_2$. Let $K_2 \subseteq W_2 \subseteq \overline{W}_2 \subseteq V_2$ $\overline{W}_2$ compact. Continue this way finally obtaining $W_1, \cdots, W_n$, $K \subseteq W_1 \cup \cdots \cup W_n$, and $\overline{W}_i \subseteq V_i$ $\overline{W}_i$ compact. Now let

$$\overline{W}_i \subseteq U_i \subseteq \overline{U}_i \subseteq V_i, \overline{U}_i$$

compact.



By Lemma 3.12.4, let $\overline{U}_i \prec \phi_i \prec V_i$, $\cup_{i=1}^n \overline{W}_i \prec \gamma \prec \cup_{i=1}^n U_i$. Define

$$\psi_i(x) = \begin{cases} \gamma(x)\phi_i(x)/\sum_{j=1}^n \phi_j(x) \text{ if } \sum_{j=1}^n \phi_j(x) \neq 0, \\ 0 \text{ if } \sum_{j=1}^n \phi_j(x) = 0. \end{cases}$$

If x is such that $\sum_{j=1}^n \phi_j(x) = 0$, then $x \notin \cup_{i=1}^n \overline{U}_i$. Consequently $\gamma(y) = 0$ for all y near x and so $\psi_i(y) = 0$ for all y near x. Hence ψ_i is continuous at such x. If $\sum_{j=1}^n \phi_j(x) \neq 0$, this situation persists near x and so ψ_i is continuous at such points. Therefore ψ_i is continuous. If $x \in K$, then $\gamma(x) = 1$ and so $\sum_{j=1}^n \psi_j(x) = 1$. Clearly $0 \leq \psi_i(x) \leq 1$ and $\text{spt}(\psi_j) \subseteq V_j$. ■

3.13 Exercises

1. Let $d(x,y) = |x-y|$ for $x, y \in \mathbb{R}$. Show that this is a metric on $\mathbb{R}$.

2. Now consider $\mathbb{R}^n$. Let $\|\boldsymbol{x}\|_\infty \equiv \max\{|x_i|, i = 1, \cdots, n\}$. Define $d(\boldsymbol{x}, \boldsymbol{y}) \equiv \|\boldsymbol{x} - \boldsymbol{y}\|_\infty$. Show that this is a metric on $\mathbb{R}^n$. In the case of $n = 2$, describe the ball $B(\mathbf{0}, r)$. **Hint:** First show that $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$.

3. Let $C([0,T])$ denote the space of functions which are continuous on $[0,T]$. Define

$$\|f\| \equiv \|f\|_\infty \equiv \sup_{t \in [0,T]} |f(t)| = \max_{t \in [0,T]} |f(t)|$$

   Verify the following. $\|f+g\| \leq \|f\| + \|g\|$. Then use to show that $d(f,g) \equiv \|f-g\|$ is a metric and that with this metric, $(C([0,T]), d)$ is a metric space.

4. Recall that $[a,b]$ is compact. Also, it is Lemma 3.5.9 above. Thus every open cover has a finite subcover of the set. Also recall that a sequence of numbers $\{x_n\}$ is a Cauchy sequence means that for every $\varepsilon > 0$ there exists N such that if $m, n > N$, then $|x_n - x_m| < \varepsilon$. First show that every Cauchy sequence is bounded. Next, using the compactness of closed intervals, show that every Cauchy sequence has a convergent subsequence. By Theorem 3.2.2, the original Cauchy sequence converges. Thus $\mathbb{R}$ with the usual metric just described is complete because every Cauchy sequence converges.

5. Using the result of the above problem, show that $(\mathbb{R}^n, \|\cdot\|_\infty)$ is a complete metric space. That is, every Cauchy sequence converges. Here $d(\boldsymbol{x}, \boldsymbol{y}) \equiv \|\boldsymbol{x} - \boldsymbol{y}\|_\infty$.

6. Suppose you had (X_i, d_i) is a metric space. Now consider the product space $X \equiv \prod_{i=1}^n X_i$ with $d(\boldsymbol{x}, \boldsymbol{y}) = \max\{d(x_i, y_i), i = 1 \cdots, n\}$. Would this be a metric space? If so, prove that this is the case.

Does triangle inequality hold? **Hint:** For each i,

$$d_i(x_i,z_i) \leq d_i(x_i,y_i) + d_i(y_i,z_i) \leq d(\boldsymbol{x},\boldsymbol{y}) + d(\boldsymbol{y},\boldsymbol{z})$$

Now take max of the two ends.

7. In the above example, if each (X_i,d_i) is complete, explain why (X,d) is also complete.

8. Show that $C([0,T])$ is a complete metric space. That is, show that if $\{f_n\}$ is a Cauchy sequence, then there exists $f \in C([0,T])$ such that

$$\lim_{n\to\infty} d(f,f_n) = \lim_{n\to\infty} \|f - f_n\| = 0$$

This is just a special case of theorems discussed in the chapter.

9. Let X be a nonempty set of points. Say it has infinitely many points. Define $d(x,y) = 1$ if $x \neq y$ and $d(x,y) = 0$ if $x = y$. Show that this is a metric. Show that in (X,d) every point is open and closed. In fact, show that every set is open and every set is closed. Is this a complete metric space? Explain why. Describe the open balls.

10. Show that the union of any set of open sets is an open set. Show the intersection of any set of closed sets is closed. Let A be a nonempty subset of a metric space (X,d). Then the closure of A, written as $\bar{A}$ is defined to be the intersection of all closed sets which contain A. Show that $\bar{A} = A \cup A'$. That is, to find the closure, you just take the set and include all limit points of the set. It was proved in the chapter, but go over it yourself.

11. Let A' denote the set of limit points of A, a nonempty subset of a metric space (X,d). Show that A' is closed.

12. A theorem was proved which gave three equivalent descriptions of compactness of a metric space. One of them said the following: A metric space is compact if and only if it is complete and totally bounded. Suppose (X,d) is a complete metric space and $K \subseteq X$. Then (K,d) is also clearly a metric space having the same metric as X. Show that (K,d) is compact if and only if it is **closed** and totally bounded. Note the similarity with the Heine Borel theorem on $\mathbb{R}$. Show that on $\mathbb{R}$, every bounded set is also totally bounded. Thus the earlier Heine Borel theorem for $\mathbb{R}$ is obtained.

13. Suppose (X_i,d_i) is a compact metric space. Then the Cartesian product is also a metric space. That is $(\prod_{i=1}^n X_i, d)$ is a metric space where $d(\boldsymbol{x},\boldsymbol{y}) \equiv \max\{d_i(x_i,y_i)\}$. Show that $(\prod_{i=1}^n X_i, d)$ is compact. Recall the Heine Borel theorem for $\mathbb{R}$. Explain why $\prod_{i=1}^n [a_i,b_i]$ is compact in $\mathbb{R}^n$ with the distance given by

$$d(\boldsymbol{x},\boldsymbol{y}) = \max\{|x_i - y_i|\}$$

Hint: It suffices to show that $(\prod_{i=1}^n X_i, d)$ is sequentially compact. Let $\{\boldsymbol{x}^m\}_{m=1}^\infty$ be a sequence. Then $\{x_1^m\}_{m=1}^\infty$ is a sequence in X_i. Therefore, it has a subsequence $\left\{x_1^{k_1}\right\}_{k_1=1}^\infty$ which converges to a point $x_1 \in X_1$. Now consider $\left\{x_2^{k_1}\right\}_{k_1=1}^\infty$ the second components. It has a subsequence denoted as k_2 such that $\left\{x_2^{k_2}\right\}_{k_2=1}^\infty$ converges to a

point x_2 in X_2. Explain why $\lim_{k_2\to\infty} x_1^{k_2} = x_1$. Continue doing this n times. Explain why $\lim_{k_n\to\infty} x_l^{k_n} = x_l \in X_l$ for each l. Then explain why this is the same as saying $\lim_{k_n\to\infty} \boldsymbol{x}^{k_n} = \boldsymbol{x}$ in $\left(\prod_{i=1}^n X_i, d\right)$.

14. If you have a metric space (X,d) and a compact subset of (X,d) K, suppose that L is a closed subset of K. Explain why L must also be compact. **Hint:** Use the definition of compactness. Explain why every closed and bounded set in $\mathbb{R}^n$ is compact. Here the distance is given by $d(\boldsymbol{x},\boldsymbol{y}) \equiv \max_{1\le i\le n}\{|x_i - y_i|\}$.

15. Show that compactness is a topological property. If $(X,d),(Y,\rho)$ are both metric spaces and $f: X \to Y$ has the property that f is one to one, onto, and continuous, and also f^{-1} is one to one onto and continuous, then the two metric spaces are compact or not compact together. That is one is compact if and only if the other is.

16. Consider $\mathbb{R}$ the real numbers. Define a distance in the following way. $\rho(x,y) \equiv |\arctan(x) - \arctan(y)|$ Show this is a good enough distance and that the open sets which come from this distance are the same as the open sets which come from the usual distance $d(x,y) = |x-y|$. Explain why this yields that the identity mapping $f(x) = x$ is continuous with continuous inverse as a map from $(\mathbb{R},d)$ to $(\mathbb{R},\rho)$. To do this, you show that an open ball taken with respect to one of these is also open with respect to the other. However, $(\mathbb{R},\rho)$ is not a complete metric space while $(\mathbb{R},d)$ is. Thus, unlike compactness. Completeness is not a topological property. **Hint:** To show the lack of completeness of $(\mathbb{R},\rho)$, consider $x_n = n$. Show it is a Cauchy sequence with respect to ρ.

17. If K is a compact subset of (X,d) and $y \notin K$, show that there always exists $x \in K$ such that $d(x,y) = \operatorname{dist}(y,K)$. Give an example in $\mathbb{R}$ to show that this might not be so if K is not compact.

18. If S is a nonempty set, the diameter of S denoted as $\operatorname{diam}(S)$ is defined as follows. $\operatorname{diam}(S) \equiv \sup\{d(x,y) : x,y \in S\}$. Suppose (X,d) is a complete metric space and you have a nested sequence of closed sets whose diameters converge to 0. That is, each A_n is closed, $\cdots A_n \supseteq A_{n+1} \cdots$ and $\lim_{n\to\infty} \operatorname{diam}(A_n) = 0$. Show that there is exactly one point p contained in the intersection of all these sets A_n. Give an example which shows that if the condition on the diameters does not hold, then maybe there is no point in the intersection of these sets.

19. Two metric spaces $(X,d),(Y,\rho)$ are homeomorphic if there exists a continuous function $f: X \to Y$ which is one to one onto, and whose inverse is also continuous one to one and onto. Show that the interval $[0,1]$ is not homeomorphic to the unit circle. **Hint:** Recall that the continuous image of a connected set is connected, Theorem 3.11.3. However, if you remove a point from $[0,1]$ it is no longer connected but removing a single point from the circle results in a connected set.

20. Using the same methods in the above problem, show that the unit circle is not homeomorphic to the unit sphere $\left\{x^2+y^2+z^2 = 1\right\}$ and the unit circle is not homeomorphic to a figure eight.

21. The rational numbers $\mathbb{Q}$ and the natural numbers $\mathbb{N}$ have the property that there is a one to one and onto map from $\mathbb{N}$ to $\mathbb{Q}$. This is a simple consequence of the Schroeder Bernstein theorem presented earlier. Both of these are also metric spaces with respect

to the usual metric on $\mathbb{R}$. Are they homeomorphic? **Hint:** Suppose they were. Then in $\mathbb{Q}$ consider $(1,2)$, all the rationals between 1 and 2 excluding 1 and 2. This is not a closed set because 2 is a limit point of the set which is not in it. Now if you have f a homeomorphism, consider $f((1,2))$. Is this set closed?

22. If you have an open set O in $\mathbb{R}$, show that O is the countable union of disjoint open intervals. **Hint:** Consider the connected components. Go over this for yourself. It is in the chapter.

23. Addition and multiplication on $\mathbb{R}$ can be considered mappings from $\mathbb{R}\times\mathbb{R}$ to $\mathbb{R}$ as follows. $+(x,y)\equiv x+y, \cdot(x,y)\equiv xy$. Here the metric on $\mathbb{R}\times\mathbb{R}$ can be taken as $d((x,y),(\hat{x},\hat{y}))\equiv \max(|x-\hat{x}|,|y-\hat{y}|)$. Show these operations are continuous functions.

24. Suppose K is a compact subset of a metric space (X,d) and there is an open cover $\mathscr{C}$ of K. Show that there exists a single positive $\delta>0$ such that if $x\in K, B(x,\delta)$ is contained in some set of $\mathscr{C}$. This number is called a Lebesgue number. Do this directly from the definition of compactness in terms of open covers without using the equivalence of compactness and sequential compactness.

25. Show uniform continuity of a continuous function defined on a compact set where compactness only refers to open covers. Use the above problem on existence of the Lebesgue number.

26. Let $f:D\to\mathbb{R}$ be a function. This function is said to be lower semicontinuous[1] at $x\in D$ if for any sequence $\{x_n\}\subseteq D$ which converges to x it follows $f(x)\leq \liminf_{n\to\infty} f(x_n)$. Suppose D is sequentially compact and f is lower semicontinuous at every point of D. Show that then f achieves its minimum on D. Here D is some metric space. Let $f:D\to\mathbb{R}$ be a function. This function is said to be upper semicontinuous at $x\in D$ if for any sequence $\{x_n\}\subseteq D$ which converges to x it follows $f(x)\geq \limsup_{n\to\infty} f(x_n)$. Suppose D is sequentially compact and f is upper semicontinuous at every point of D. Show that then f achieves its maximum on D.

27. Show that a real valued function defined on a metric space D is continuous if and only if it is both upper and lower semicontinuous.

28. Give an example of a lower semicontinuous function defined on $\mathbb{R}$ which is not continuous and an example of an upper semicontinuous function which is not continuous.

29. More generally, one considers functions which have values in $[-\infty,\infty]$. Then f is upper semicontinuous if, whenever $x_n\to x, f(x)\geq \limsup_{n\to\infty} f(x_n)$ and lower semicontinuous if whenever $x_n\to x,\ f(x)\leq \liminf_{n\to\infty} f(x_n)$. Suppose $\{f_\alpha:\alpha\in\Lambda\}$ is a collection of continuous real valued functions defined on a metric space. Let $F(x)\equiv \inf\{f_\alpha(x):\alpha\in\Lambda\}$. Show F is an upper semicontinuous function. Next let $G(x)\equiv \sup\{f_\alpha(x):\alpha\in\Lambda\}$. Show G is a lower semicontinuous function.

30. The result of this problem is due to Hausdorff. It says that if you have any lower semicontinuous real valued function defined on a metric space (X,d), then it is the

[1] The notion of lower semicontinuity is very important for functions which are defined on infinite dimensional sets.

limit of an increasing sequence of continuous functions. Here is an outline. You complete the details.

(a) First suppose $f(x) \geq 0$ for all x. Define $f_n(x) \equiv \inf_{z \in X} \{f(z) + nd(z,x)\}$. Then $f(x) \geq f_n(x)$ and $f_n(x)$ is increasing in n. Also each f_n is continuous because $f_n(x) \leq f(z) + nd(z,y) + nd(y,x)$. Thus $f_n(x) \leq f_n(y) + nd(y,x)$. Why? It follows that $|f_n(x) - f_n(y)| \leq nd(y,x)$. Why?

(b) Let $h(x) = \lim_{n\to\infty} f_n(x)$. Then $h(x) \leq f(x)$. Why? Now for each $\varepsilon > 0$, and fixed x, there exists z_n such that $f_n(x) + \varepsilon > f(z_n) + nd(z_n,x)$Why? Therefore, $z_n \to x$. Why?

(c) Then

$$\begin{aligned} h(x) + \varepsilon &= \lim_{n\to\infty} f_n(x) + \varepsilon \geq \lim\inf_{n\to\infty} (f(z_n) + nd(z_n,x)) \\ &\geq \lim\inf_{n\to\infty} f(z_n) \geq f(x) \end{aligned}$$

Why? Therefore, $h(x) \geq f(x)$ and so they are equal. Why?

(d) Now consider $f : X \to (-\infty, \infty)$ and is lower semicontinuous as just explained. Consider $\frac{\pi}{2} + \arctan f(x) \equiv g(x)$. Then $\arctan f(x) \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ because f has real values. Then $g(x)$ is also lower semicontinuous having values in $(0, \pi)$. Why? By what was just shown, there exists $g_n(x) \uparrow g(x)$ where each g_n is continuous. Consider $f_n(x) \equiv \tan\left(g_n(x) - \frac{\pi}{2}\right)$. Then f_n is continuous and increases to $f(x)$.

31. Generalize the above problem to the case where f is an upper semicontinuous real valued function. That is, $f(x) \geq \limsup_{n\to\infty} f(x_n)$ whenever $x_n \to x$. Show there are continuous functions $\{f_n(x)\}$ such that $f_n(x) \downarrow f(x)$. **Hint** To save trouble, maybe show that f is upper semicontinuous if and only if $-f$ is lower semicontinuous. Then maybe you could just use the above problem.

32. What if f is lower (upper) semicontinuous with values in $[-\infty, \infty]$? In this case, you consider $[-\infty, \infty]$ as a metric space as follows:$d(x,y) \equiv |\arctan(x) - \arctan(y)|$. Then you can generalize the above problems to show that if f is lower semicontinuous with values into $[-\infty, \infty]$ then it is the increasing limit of continuous functions with values in $[-\infty, \infty]$. Note that in this case a function identically equal to ∞ would be continuous so this is a rather odd sort of thing, a little different from what we normally like to consider. Check the details and explain why in this setting, the lower semicontinuous functions are exactly pointwise limits of increasing sequences of continuous functions and the upper semicontinuous functions are exactly pointwise limits of decreasing sequences of continuous functions.

33. This is a nice result in Taylor [56]. For a nonempty set $T, \partial T$ is the set of points p such that $B(p,r)$ contains points of T and points of T^C for each $r > 0$. Suppose you have T a proper subset of a metric space and S is a connected, nonempty set such that $S \cap T \neq \emptyset, S \cap T^C \neq \emptyset$. Show that S must contain a point of ∂T.

Chapter 4

Linear Spaces

The thing which is missing in the above material about metric spaces is any kind of algebra. In most applications, we are interested in adding things and multiplying things by scalars and so forth. This requires the notion of a vector space, also called a linear space. The simplest example is $\mathbb{R}^n$ which is described next.

In this chapter, $\mathbb{F}$ will refer to either $\mathbb{R}$ or $\mathbb{C}$. It doesn't make any difference to the arguments which it is and so $\mathbb{F}$ is written to symbolize whichever you wish to think about. When it is desired to emphasize that certain quantities are vectors, bold face will often be used. This is not necessarily done consistently. Sometimes context is considered sufficient.

4.1 Algebra in $\mathbb{F}^n$, Vector Spaces

There are exactly two algebraic operations done with elements of $\mathbb{F}^n$. One is addition and the other is multiplication by numbers, called scalars. In the case of $\mathbb{C}^n$ the scalars are complex numbers while in the case of $\mathbb{R}^n$ the only allowed scalars are real numbers. Thus, the scalars always come from $\mathbb{F}$ in either case.

Definition 4.1.1 *If $\boldsymbol{x} \in \mathbb{F}^n$ and $a \in \mathbb{F}$, also called a scalar, then $a\boldsymbol{x} \in \mathbb{F}^n$ is defined by*

$$a\boldsymbol{x} = a(x_1, \cdots, x_n) \equiv (ax_1, \cdots, ax_n). \tag{4.1}$$

This is known as scalar multiplication. If $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{F}^n$ then $\boldsymbol{x} + \boldsymbol{y} \in \mathbb{F}^n$ and is defined by

$$\begin{aligned} \boldsymbol{x} + \boldsymbol{y} &= (x_1, \cdots, x_n) + (y_1, \cdots, y_n) \\ &\equiv (x_1 + y_1, \cdots, x_n + y_n) \end{aligned} \tag{4.2}$$

the points in $\mathbb{F}^n$ are also referred to as vectors.

Actually, in dealing with vectors in $\mathbb{F}^n$, it is more customary in linear algebra to write them as column vectors. To save space, I will sometimes write $(x_1, \cdots, x_n)^T$ to indicate the column vector having x_1 on the top and x_n on the bottom. With this definition, the algebraic properties satisfy the conclusions of the following theorem. These conclusions are called the vector space axioms. Any time you have a set and a field of scalars satisfying the axioms of the following theorem, it is called a vector space or linear space.

Theorem 4.1.2 *For $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{F}^n$ and α, β scalars, (real numbers), the following hold.*

$$\boldsymbol{v} + \boldsymbol{w} = \boldsymbol{w} + \boldsymbol{v}, \tag{4.3}$$

the commutative law of addition,

$$(\boldsymbol{v} + \boldsymbol{w}) + \boldsymbol{z} = \boldsymbol{v} + (\boldsymbol{w} + \boldsymbol{z}), \tag{4.4}$$

the associative law for addition,

$$\boldsymbol{v} + \boldsymbol{0} = \boldsymbol{v}, \tag{4.5}$$

the existence of an additive identity,

$$\boldsymbol{v} + (-\boldsymbol{v}) = \boldsymbol{0}, \tag{4.6}$$

the existence of an additive inverse, Also

$$\alpha(\boldsymbol{v}+\boldsymbol{w}) = \alpha\boldsymbol{v}+\alpha\boldsymbol{w}, \tag{4.7}$$

$$(\alpha+\beta)\boldsymbol{v} = \alpha\boldsymbol{v}+\beta\boldsymbol{v}, \tag{4.8}$$

$$\alpha(\beta\boldsymbol{v}) = \alpha\beta(\boldsymbol{v}), \tag{4.9}$$

$$1\boldsymbol{v} = \boldsymbol{v}. \tag{4.10}$$

In the above $\mathbf{0} = (0,\cdots,0)$.

You should verify these properties all hold. For example, consider 4.7

$$\begin{aligned}\alpha(\boldsymbol{v}+\boldsymbol{w}) &= \alpha(v_1+w_1,\cdots,v_n+w_n)\\ &= (\alpha(v_1+w_1),\cdots,\alpha(v_n+w_n))\\ &= (\alpha v_1+\alpha w_1,\cdots,\alpha v_n+\alpha w_n)\\ &= (\alpha v_1,\cdots,\alpha v_n)+(\alpha w_1,\cdots,\alpha w_n)\\ &= \alpha\boldsymbol{v}+\alpha\boldsymbol{w}.\end{aligned}$$

As usual subtraction is defined as $\boldsymbol{x}-\boldsymbol{y} \equiv \boldsymbol{x}+(-\boldsymbol{y})$.

4.2 Subspaces Spans and Bases

As mentioned above, $\mathbb{F}^n$ is an example of a vector space. In dealing with vector spaces, the concept of linear combination is fundamental. When one considers only algebraic considerations, it makes no difference what field of scalars you are using. It could be $\mathbb{R}$, $\mathbb{C}$, $\mathbb{Q}$ or even a field of residue classes. However, go ahead and think $\mathbb{R}$ or $\mathbb{C}$ since the subject of interest here is analysis.

Definition 4.2.1 *Let* $\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_p\}$ *be vectors in a vector space* Y *having the field of scalars* $\mathbb{F}$. *A linear combination is any expression of the form* $\sum_{i=1}^{p} c_i\boldsymbol{x}_i$ *where the* c_i *are scalars. The set of all linear combinations of these vectors is called* $\operatorname{span}(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_p)$. *A vector* $\boldsymbol{v}$ *is said to be in the span of some set* S *of vectors if* $\boldsymbol{v}$ *is a linear combination of vectors of* S. ***This means: finite linear combination.*** *If* $V \subseteq Y$, *then* V *is called a subspace if it contains* $\mathbf{0}$ *and whenever* α,β *are scalars and* $\boldsymbol{u}$ *and* $\boldsymbol{v}$ *are vectors of* V, *it follows* $\alpha\boldsymbol{u}+\beta\boldsymbol{v} \in V$. *That is, it is "closed under the algebraic operations of vector addition and scalar multiplication" and is therefore, a vector space. A linear combination of vectors is said to be trivial if all the scalars in the linear combination equal zero. A set of vectors is said to be linearly independent if the only linear combination of these vectors which equals the zero vector is the trivial linear combination. Thus* $\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\}$ *is called linearly independent if whenever* $\sum_{k=1}^{n} c_k\boldsymbol{x}_k = \mathbf{0}$, *it follows that all the scalars,* c_k *equal zero. A set of vectors,* $\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\}$, *is called linearly dependent if it is not linearly independent. Thus the set of vectors is linearly dependent if there exist scalars,* $c_i, i = 1,\cdots,n$, *not all zero such that* $\sum_{k=1}^{n} c_k\boldsymbol{x}_k = \mathbf{0}$.

Lemma 4.2.2 *A set of vectors* $\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\}$ *is linearly independent if and only if none of the vectors can be obtained as a linear combination of the others.*

Proof: Suppose first that $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ is linearly independent. If

$$\boldsymbol{x}_k = \sum_{j \neq k} c_j \boldsymbol{x}_j,$$

then $\mathbf{0} = 1\boldsymbol{x}_k + \sum_{j \neq k} (-c_j) \boldsymbol{x}_j$, a nontrivial linear combination, contrary to assumption. This shows that if the set is linearly independent, then none of the vectors is a linear combination of the others.

Now suppose no vector is a linear combination of the others. Is $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ linearly independent? If it is not, there exist scalars, c_i, not all zero such that $\sum_{i=1}^n c_i \boldsymbol{x}_i = \mathbf{0}$. Say $c_k \neq 0$. Then you can solve for $\boldsymbol{x}_k$ as $\boldsymbol{x}_k = \sum_{j \neq k} (-c_j/c_k) \boldsymbol{x}_j$ contrary to assumption. This proves the lemma. ■

The following is called the exchange theorem.

Theorem 4.2.3 *If*

$$\text{span}(\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r) \subseteq \text{span}(\boldsymbol{v}_1, \cdots, \boldsymbol{v}_s) \equiv V$$

and $\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r\}$ *are linearly independent, then* $r \leq s$.

Proof: Suppose $r > s$. Let F_p denote the first p vectors in $\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r\}$. Let F_0 denote the empty set. Let E_p denote a finite list of vectors of $\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_s\}$ and let $|E_p|$ denote the number of vectors in the list. Note that, by assumption, $\text{span}(F_0, E_s) = V$. For $0 \leq p \leq s$, let E_p have the property $\text{span}(F_p, E_p) = V$ and $|E_p|$ is as small as possible for this to happen. If $|E_p| = 0$, then $\text{span}(F_p) = V$ which would imply that, since $r > s \geq p, \boldsymbol{u}_r \in \text{span}(F_s)$ contradicting the linear independence of $\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r\}$. Assume then that $|E_p| > 0$. Then $\boldsymbol{u}_{p+1} \in \text{span}(F_p, E_p)$ and so there are constants, $c_1, \cdots, c_p$ and $d_1, \cdots, d_m$ such that $\boldsymbol{u}_{p+1} = \sum_{i=1}^p c_i \boldsymbol{u}_i + \sum_{j=1}^m d_i \boldsymbol{z}_j$ for $\{\boldsymbol{z}_1, \cdots, \boldsymbol{z}_m\} \subseteq \{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_s\}$. Then not all the d_i can equal zero because this would violate the linear independence of the $\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r\}$. Therefore, you can solve for one of the $\boldsymbol{z}_k$ as a linear combination of $\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_{p+1}\}$ and the other $\boldsymbol{z}_j$. Thus you can change F_p to F_{p+1} and include one fewer vector in E_{p+1} with $\text{span}(F_{p+1}, E_{p+1}) = V$ and so $|E_{p+1}| < |E_p|$ contrary to the claim that $|E_p|$ was as small as possible. Thus $|E_p| = 0$ after all and so a contradiction results.

Alternate proof: Recall from linear algebra that if you have A an $m \times n$ matrix where $m < n$ so there are more columns than rows, then there exists a nonzero solution $\boldsymbol{x}$ to the equation $A\boldsymbol{x} = \mathbf{0}$. Recall why this was. You must have free variables. Then by assumption, you have $\boldsymbol{u}_j = \sum_{i=1}^s a_{ij} \boldsymbol{v}_i$. If $s < r$, then the matrix (a_{ij}) has more columns than rows and so there exists a nonzero vector $\boldsymbol{x} \in \mathbb{F}^r$ such that $\sum_{j=1}^r a_{ij} x_j = 0$. Then consider the following.

$$\sum_{j=1}^r x_j \boldsymbol{u}_j = \sum_{j=1}^r x_j \sum_{i=1}^s a_{ij} \boldsymbol{v}_i = \sum_i \sum_j a_{ij} x_j \boldsymbol{v}_i = \sum_i 0 \boldsymbol{v}_j = \mathbf{0}$$

and since not all $x_j = 0$, this contradicts the independence of $\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r\}$. ■

Definition 4.2.4 *A finite set of vectors,* $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_r\}$ *is a basis for a vector space V if*

$$\text{span}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_r) = V$$

and $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_r\}$ *is linearly independent. Thus if* $\boldsymbol{v} \in V$ *there exist unique scalars,* $v_1, \cdots, v_r$ *such that* $\boldsymbol{v} = \sum_{i=1}^r v_i \boldsymbol{x}_i$. *These scalars are called the components of* $\boldsymbol{v}$ *with respect to the basis* $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_r\}$ *and* $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_r\}$ *are said to "span" V.*

Corollary 4.2.5 *Let $\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_r\}$ and $\{\boldsymbol{y}_1,\cdots,\boldsymbol{y}_s\}$ be two bases[1] of $\mathbb{F}^n$. Then $r=s=n$. More generally, if you have two bases for a vector space V then they have the same number of vectors.*

Proof: From the exchange theorem, Theorem 4.2.3, if

$$\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_r\},\{\boldsymbol{y}_1,\cdots,\boldsymbol{y}_s\}$$

are two bases for V, then $r\leq s$ and $s\leq r$. Now note the vectors,

$$\boldsymbol{e}_i=\overbrace{(0,\cdots,0,1,0\cdots,0)^T}^{1 \text{ is in the } i^{th} \text{ slot}}$$

for $i=1,2,\cdots,n$ are a basis for $\mathbb{F}^n$. ■

Lemma 4.2.6 *Let $\{\boldsymbol{v}_1,\cdots,\boldsymbol{v}_r\}$ be a set of vectors. Then $V\equiv \operatorname{span}(\boldsymbol{v}_1,\cdots,\boldsymbol{v}_r)$ is a subspace.*

Proof: Suppose α,β are two scalars and let $\sum_{k=1}^r c_k\boldsymbol{v}_k$ and $\sum_{k=1}^r d_k\boldsymbol{v}_k$ are two elements of V. What about $\alpha\sum_{k=1}^r c_k\boldsymbol{v}_k+\beta\sum_{k=1}^r d_k\boldsymbol{v}_k$? Is it also in V?

$$\alpha\sum_{k=1}^r c_k\boldsymbol{v}_k+\beta\sum_{k=1}^r d_k\boldsymbol{v}_k=\sum_{k=1}^r(\alpha c_k+\beta d_k)\boldsymbol{v}_k\in V$$

so the answer is yes. It is clear that 0 is in $\operatorname{span}(\boldsymbol{v}_1,\cdots,\boldsymbol{v}_r)$. This proves the lemma. ■

Definition 4.2.7 *Let V be a vector space. It is finite dimensional when it has a basis of finitely many vectors. Otherwise, it is infinite dimensional. Then $\dim(V)$ read as the dimension of V is the number of vectors in a basis.*

Of course you should wonder right now whether an arbitrary subspace of a finite dimensional vector space even has a basis. In fact it does and this is in the next theorem. First, here is an interesting lemma.

Lemma 4.2.8 *Suppose $\boldsymbol{v}\notin\operatorname{span}(\boldsymbol{u}_1,\cdots,\boldsymbol{u}_k)$ and $\{\boldsymbol{u}_1,\cdots,\boldsymbol{u}_k\}$ is linearly independent. Then $\{\boldsymbol{u}_1,\cdots,\boldsymbol{u}_k,\boldsymbol{v}\}$ is also linearly independent.*

Proof: Suppose $\sum_{i=1}^k c_i\boldsymbol{u}_i+d\boldsymbol{v}=\boldsymbol{0}$. It is required to verify that each $c_i=0$ and that $d=0$. But if $d\neq 0$, then you can solve for $\boldsymbol{v}$ as a linear combination of the vectors, $\{\boldsymbol{u}_1,\cdots,\boldsymbol{u}_k\}$, $\boldsymbol{v}=-\sum_{i=1}^k\left(\frac{c_i}{d}\right)\boldsymbol{u}_i$ contrary to assumption. Therefore, $d=0$. But then $\sum_{i=1}^k c_i\boldsymbol{u}_i=0$ and the linear independence of $\{\boldsymbol{u}_1,\cdots,\boldsymbol{u}_k\}$ implies each $c_i=0$ also. ■

Theorem 4.2.9 *Let V be a nonzero subspace of Y a finite dimensional vector space having dimension n. Then V has a basis.*

[1] This is the plural form of basis. We could say basiss but it would involve an inordinate amount of hissing as in "The sixth shiek's sixth sheep is sick". This is the reason that bases is used instead of basiss.

Proof: Let $\boldsymbol{v}_1 \in V$ where $\boldsymbol{v}_1 \neq 0$. If $\text{span}\{\boldsymbol{v}_1\} = V$, stop. $\{\boldsymbol{v}_1\}$ is a basis for V. Otherwise, there exists $\boldsymbol{v}_2 \in V$ which is not in $\text{span}\{\boldsymbol{v}_1\}$. By Lemma 4.2.8 $\{\boldsymbol{v}_1, \boldsymbol{v}_2\}$ is a linearly independent set of vectors. If $\text{span}\{\boldsymbol{v}_1, \boldsymbol{v}_2\} = V$ stop, $\{\boldsymbol{v}_1, \boldsymbol{v}_2\}$ is a basis for V. If $\text{span}\{\boldsymbol{v}_1, \boldsymbol{v}_2\} \neq V$, then there exists $\boldsymbol{v}_3 \notin \text{span}\{\boldsymbol{v}_1, \boldsymbol{v}_2\}$ and $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3\}$ is a larger linearly independent set of vectors. Continuing this way, the process must stop before $n+1$ steps because if not, it would be possible to obtain $n+1$ linearly independent vectors contrary to the exchange theorem, Theorem 4.2.3, and the assumed dimension of Y. ■

In words the following corollary states that any linearly independent set of vectors can be enlarged to form a basis.

Corollary 4.2.10 *Let V be a subspace of Y, a finite dimensional vector space of dimension n and let $\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_r\}$ be a linearly independent set of vectors in V. Then either it is a basis for V or there exist vectors, $\boldsymbol{v}_{r+1}, \cdots, \boldsymbol{v}_s$ such that*

$$\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_r, \boldsymbol{v}_{r+1}, \cdots, \boldsymbol{v}_s\}$$

is a basis for V.

Proof: This follows immediately from the proof of Theorem 4.2.9. You do exactly the same argument except you start with $\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_r\}$ rather than $\{\boldsymbol{v}_1\}$. ■

It is also true that any spanning set of vectors can be restricted to obtain a basis.

Theorem 4.2.11 *Let V be a subspace of Y, a finite dimensional vector space of dimension n and suppose $\text{span}(\boldsymbol{u}_1 \cdots, \boldsymbol{u}_p) = V$ where the $\boldsymbol{u}_i$ are nonzero vectors. Then there exist vectors, $\{\boldsymbol{v}_1 \cdots, \boldsymbol{v}_r\}$ such that $\{\boldsymbol{v}_1 \cdots, \boldsymbol{v}_r\} \subseteq \{\boldsymbol{u}_1 \cdots, \boldsymbol{u}_p\}$ and $\{\boldsymbol{v}_1 \cdots, \boldsymbol{v}_r\}$ is a basis for V.*

Proof: Let r be the smallest positive integer with the property that for some set,

$$\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_r\} \subseteq \{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_p\}, \text{span}(\boldsymbol{v}_1, \cdots, \boldsymbol{v}_r) = V.$$

Then $r \leq p$ and it must be the case that $\{\boldsymbol{v}_1 \cdots, \boldsymbol{v}_r\}$ is linearly independent because if it were not so, one of the vectors, say $\boldsymbol{v}_k$ would be a linear combination of the others. But then you could delete this vector from $\{\boldsymbol{v}_1 \cdots, \boldsymbol{v}_r\}$ and the resulting list of $r-1$ vectors would still span V contrary to the definition of r. ■

4.3 Inner Product and Normed Linear Spaces

4.3.1 The Inner Product in $\mathbb{F}^n$

To do calculus, you must understand what you mean by distance. For functions of one variable, the distance was provided by the absolute value of the difference of two numbers. This must be generalized to $\mathbb{F}^n$ and to more general situations.

Definition 4.3.1 *Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{F}^n$. Thus $\boldsymbol{x} = (x_1, \cdots, x_n)$ where each $x_k \in \mathbb{F}$ and a similar formula holding for $\boldsymbol{y}$. Then the inner product of these two vectors is defined to be*

$$(\boldsymbol{x}, \boldsymbol{y}) \equiv \sum_j x_j \overline{y_j} \equiv x_1 \overline{y_1} + \cdots + x_n \overline{y_n}.$$

Sometimes it is denoted as $\boldsymbol{x} \cdot \boldsymbol{y}$.

Notice how you put the conjugate on the entries of the vector $\boldsymbol{y}$. It makes no difference if the vectors happen to be real vectors but with complex vectors you must involve a conjugate. The reason for this is that when you take the inner product of a vector with itself, you want to get the square of the length of the vector, a positive number. Placing the conjugate on the components of $\boldsymbol{y}$ in the above definition assures this will take place. Thus $(\boldsymbol{x},\boldsymbol{x}) = \sum_j x_j \overline{x_j} = \sum_j |x_j|^2 \geq 0$. If you didn't place a conjugate as in the above definition, things wouldn't work out correctly. For example, $(1+i)^2 + 2^2 = 4+2i$ and this is not a positive number.

The following properties of the inner product follow immediately from the definition and you should verify each of them.

Properties of the inner product:

1. $(\boldsymbol{u},\boldsymbol{v}) = \overline{(\boldsymbol{v},\boldsymbol{u})}$
2. If a,b are numbers and $\boldsymbol{u},\boldsymbol{v},\boldsymbol{z}$ are vectors then $((a\boldsymbol{u}+b\boldsymbol{v}),\boldsymbol{z}) = a(\boldsymbol{u},\boldsymbol{z}) + b(\boldsymbol{v},\boldsymbol{z})$.
3. $(\boldsymbol{u},\boldsymbol{u}) \geq 0$ and it equals 0 if and only if $\boldsymbol{u} = \boldsymbol{0}$.

Note this implies $(\boldsymbol{x},\alpha\boldsymbol{y}) = \overline{\alpha}(\boldsymbol{x},\boldsymbol{y})$ because

$$(\boldsymbol{x},\alpha\boldsymbol{y}) = \overline{(\alpha\boldsymbol{y},\boldsymbol{x})} = \overline{\alpha(\boldsymbol{y},\boldsymbol{x})} = \overline{\alpha}(\boldsymbol{x},\boldsymbol{y})$$

The norm is defined as follows.

Definition 4.3.2 *For $\boldsymbol{x} \in \mathbb{F}^n$, $|\boldsymbol{x}| \equiv \left(\sum_{k=1}^n |x_k|^2\right)^{1/2} = (\boldsymbol{x},\boldsymbol{x})^{1/2}$.*

4.3.2 General Inner Product Spaces

Any time you have a vector space which possesses an inner product, something satisfying the properties 1 - 3 above, it is called an inner product space.

Here is a fundamental inequality called the **Cauchy Schwarz inequality** which holds in any inner product space. First here is a simple lemma.

Lemma 4.3.3 *If $z \in \mathbb{F}$ there exists $\theta \in \mathbb{F}$ such that $\theta z = |z|$ and $|\theta| = 1$.*

Proof: Let $\theta = 1$ if $z = 0$ and otherwise, let $\theta = \dfrac{\overline{z}}{|z|}$. Recall that for $z = x+iy, \overline{z} = x-iy$ and $\overline{z}z = |z|^2$. In case z is real, there is no change in the above. ■

Theorem 4.3.4 *(Cauchy Schwarz)Let H be an inner product space. The following inequality holds for $\boldsymbol{x}$ and $\boldsymbol{y} \in H$.*

$$|(\boldsymbol{x},\boldsymbol{y})| \leq (\boldsymbol{x},\boldsymbol{x})^{1/2}(\boldsymbol{y},\boldsymbol{y})^{1/2} \tag{4.11}$$

Equality holds in this inequality if and only if one vector is a multiple of the other.

Proof: Let $\theta \in \mathbb{F}$ such that $|\theta| = 1$ and $\theta(\boldsymbol{x},\boldsymbol{y}) = |(\boldsymbol{x},\boldsymbol{y})|$. Consider

$$p(t) \equiv \left(\boldsymbol{x} + \overline{\theta}t\boldsymbol{y}, \boldsymbol{x} + t\overline{\theta}\boldsymbol{y}\right)$$

where $t \in \mathbb{R}$. Then from the above list of properties of the inner product,

$$\begin{aligned} 0 &\leq p(t) = (\boldsymbol{x},\boldsymbol{x}) + t\theta(\boldsymbol{x},\boldsymbol{y}) + t\overline{\theta}(\boldsymbol{y},\boldsymbol{x}) + t^2(\boldsymbol{y},\boldsymbol{y}) \\ &= (\boldsymbol{x},\boldsymbol{x}) + t\theta(\boldsymbol{x},\boldsymbol{y}) + t\overline{\theta(\boldsymbol{x},\boldsymbol{y})} + t^2(\boldsymbol{y},\boldsymbol{y}) \\ &= (\boldsymbol{x},\boldsymbol{x}) + 2t\,\mathrm{Re}(\theta(\boldsymbol{x},\boldsymbol{y})) + t^2(\boldsymbol{y},\boldsymbol{y}) \\ &= (\boldsymbol{x},\boldsymbol{x}) + 2t|(\boldsymbol{x},\boldsymbol{y})| + t^2(\boldsymbol{y},\boldsymbol{y}) \end{aligned} \tag{4.12}$$

and this must hold for all $t \in \mathbb{R}$. Therefore, if $(\boldsymbol{y},\boldsymbol{y}) = 0$ it must be the case that $|(\boldsymbol{x},\boldsymbol{y})| = 0$ also since otherwise the above inequality would be violated. Therefore, in this case, $|(\boldsymbol{x},\boldsymbol{y})| \leq (\boldsymbol{x},\boldsymbol{x})^{1/2}(\boldsymbol{y},\boldsymbol{y})^{1/2}$. On the other hand, if $(\boldsymbol{y},\boldsymbol{y}) \neq 0$, then $p(t) \geq 0$ for all t means the graph of $y = p(t)$ is a parabola which opens up and it either has exactly one real zero in the case its vertex touches the t axis or it has no real zeros. From the quadratic formula this happens exactly when $4|(\boldsymbol{x},\boldsymbol{y})|^2 - 4(\boldsymbol{x},\boldsymbol{x})(\boldsymbol{y},\boldsymbol{y}) \leq 0$ which is equivalent to 4.11.

It is clear from a computation that if one vector is a scalar multiple of the other that equality holds in 4.11. Conversely, suppose equality does hold. Then this is equivalent to saying $4|(\boldsymbol{x},\boldsymbol{y})|^2 - 4(\boldsymbol{x},\boldsymbol{x})(\boldsymbol{y},\boldsymbol{y}) = 0$ and so from the quadratic formula, there exists one real zero to $p(t) = 0$. Call it t_0. Then

$$p(t_0) \equiv \left(\boldsymbol{x} + \overline{\theta}t_0\boldsymbol{y}, \boldsymbol{x} + t_0\overline{\theta}\boldsymbol{y}\right) = \left|\boldsymbol{x} + \overline{\theta}t\boldsymbol{y}\right|^2 = 0$$

and so $\boldsymbol{x} = -\overline{\theta}t_0\boldsymbol{y}$. ■

Note that in establishing the inequality, I only used part of the above properties of the inner product. It was not necessary to use the one which says that if $(\boldsymbol{x},\boldsymbol{x}) = 0$ then $\boldsymbol{x} = \boldsymbol{0}$. That was only used to consider the case of equality.

Now the length of a vector can be defined.

Definition 4.3.5 *Let $\boldsymbol{z} \in H$. Then $|\boldsymbol{z}| \equiv (\boldsymbol{z},\boldsymbol{z})^{1/2}$.*

Theorem 4.3.6 *For length defined in Definition 4.3.5, the following hold.*

$$|\boldsymbol{z}| \geq 0 \text{ and } |\boldsymbol{z}| = 0 \text{ if and only if } \boldsymbol{z} = \boldsymbol{0} \tag{4.13}$$

$$\text{If } \alpha \text{ is a scalar, } |\alpha\boldsymbol{z}| = |\alpha||\boldsymbol{z}| \tag{4.14}$$

$$|\boldsymbol{z} + \boldsymbol{w}| \leq |\boldsymbol{z}| + |\boldsymbol{w}|. \tag{4.15}$$

Proof: The first two claims are left as exercises. To establish the third,

$$\begin{aligned} |\boldsymbol{z} + \boldsymbol{w}|^2 &\equiv (\boldsymbol{z} + \boldsymbol{w}, \boldsymbol{z} + \boldsymbol{w}) \\ &= (\boldsymbol{z},\boldsymbol{z}) + (\boldsymbol{w},\boldsymbol{w}) + (\boldsymbol{w},\boldsymbol{z}) + (\boldsymbol{z},\boldsymbol{w}) \\ &= |\boldsymbol{z}|^2 + |\boldsymbol{w}|^2 + 2\,\mathrm{Re}(\boldsymbol{w},\boldsymbol{z}) \\ &\leq |\boldsymbol{z}|^2 + |\boldsymbol{w}|^2 + 2|(\boldsymbol{w},\boldsymbol{z})| \\ &\leq |\boldsymbol{z}|^2 + |\boldsymbol{w}|^2 + 2|\boldsymbol{w}||\boldsymbol{z}| = (|\boldsymbol{z}| + |\boldsymbol{w}|)^2. \end{aligned}$$

Note that in an inner product space, you can define $d(\boldsymbol{x},\boldsymbol{y}) \equiv |\boldsymbol{x} - \boldsymbol{y}|$ and this is a metric for this inner product space. This follows from the above since d satisfies the conditions for a metric,

$$d(\boldsymbol{x},\boldsymbol{y}) = d(\boldsymbol{y},\boldsymbol{x}),\ d(\boldsymbol{x},\boldsymbol{y}) \geq 0 \text{ and equals 0 if and only if } \boldsymbol{x} = \boldsymbol{y}$$

$$d(\boldsymbol{x},\boldsymbol{y}) + d(\boldsymbol{y},\boldsymbol{z}) = |\boldsymbol{x} - \boldsymbol{y}| + |\boldsymbol{y} - \boldsymbol{z}| \geq |\boldsymbol{x} - \boldsymbol{y} + \boldsymbol{y} - \boldsymbol{z}| = |\boldsymbol{x} - \boldsymbol{z}| = d(\boldsymbol{x},\boldsymbol{z}).$$

It follows that all the theory of metric spaces developed earlier applies to this situation.

4.3.3 Normed Vector Spaces

The best sort of a norm is one which comes from an inner product. However, any vector space V which has a function $\|\cdot\|$ which maps V to $[0,\infty)$ is called a normed vector space if $\|\cdot\|$ satisfies 4.13 - 4.15. That is

$$\|z\| \geq 0 \text{ and } \|z\| = 0 \text{ if and only if } z = 0 \tag{4.16}$$

$$\text{If } \alpha \text{ is a scalar, } \|\alpha z\| = |\alpha| \|z\| \tag{4.17}$$

$$\|z + w\| \leq \|z\| + \|w\|. \tag{4.18}$$

The last inequality above is called the triangle inequality. Another version of this is

$$|\|z\| - \|w\|| \leq \|z - w\| \tag{4.19}$$

To see that 4.19 holds, note $\|z\| = \|z - w + w\| \leq \|z - w\| + \|w\|$ which implies $\|z\| - \|w\| \leq \|z - w\|$ and now switching z and w, yields $\|w\| - \|z\| \leq \|z - w\|$ which implies 4.19.

Any normed vector space is a metric space, the distance given by $d(x,y) \equiv \|x - y\|$. This satisfies all the axioms of a distance. Therefore, any normed linear space is a metric space with this metric and all the theory of metric spaces applies.

Definition 4.3.7 *When X is a normed linear space which is also complete, it is called a Banach space.*

A Banach space may or may not be finite dimensional but it is always a linear space or vector space. The field of scalars will always be $\mathbb{R}$ or $\mathbb{C}$ at least in this book. More is said about Banach spaces later.

4.3.4 The p Norms

Examples of norms are the p norms on $\mathbb{C}^n$ for $p \neq 2$. These do not come from an inner product but they are norms just the same.

Definition 4.3.8 *Let $x \in \mathbb{C}^n$. Then define for $p \geq 1$,*

$$\|x\|_p \equiv \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}.$$

The following inequality is called Holder's inequality.

Proposition 4.3.9 *For $x, y \in \mathbb{C}^n$,*

$$\sum_{i=1}^{n} |x_i| \, |y_i| \leq \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \left(\sum_{i=1}^{n} |y_i|^{p'} \right)^{1/p'}$$

The proof will depend on the following lemma shown later.

Lemma 4.3.10 *If $a, b \geq 0$ and p' is defined by $\frac{1}{p} + \frac{1}{p'} = 1$, then*

$$ab \leq \frac{a^p}{p} + \frac{b^{p'}}{p'}.$$

Proof of the Proposition: If $\boldsymbol{x}$ or $\boldsymbol{y}$ equals the zero vector there is nothing to prove. Therefore, assume they are both nonzero. Let $A = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$ and $B = \left(\sum_{i=1}^n |y_i|^{p'}\right)^{1/p'}$. Then using Lemma 4.3.10,

$$\begin{aligned}\sum_{i=1}^n \frac{|x_i|}{A}\frac{|y_i|}{B} &\leq \sum_{i=1}^n \left[\frac{1}{p}\left(\frac{|x_i|}{A}\right)^p + \frac{1}{p'}\left(\frac{|y_i|}{B}\right)^{p'}\right] \\ &= \frac{1}{p}\frac{1}{A^p}\sum_{i=1}^n |x_i|^p + \frac{1}{p'}\frac{1}{B^p}\sum_{i=1}^n |y_i|^{p'} \\ &= \frac{1}{p}+\frac{1}{p'} = 1\end{aligned}$$

and so $\sum_{i=1}^n |x_i|\,|y_i| \leq AB = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}\left(\sum_{i=1}^n |y_i|^{p'}\right)^{1/p'}$. ■

Theorem 4.3.11 *The p norms do indeed satisfy the axioms of a norm.*

Proof: It is obvious that $\|\cdot\|_p$ does indeed satisfy most of the norm axioms. The only one that is not clear is the triangle inequality. To save notation write $\|\cdot\|$ in place of $\|\cdot\|_p$ in what follows. Note also that $\frac{p}{p'} = p-1$. Then using the Holder inequality,

$$\begin{aligned}\|\boldsymbol{x}+\boldsymbol{y}\|^p &= \sum_{i=1}^n |x_i+y_i|^p \leq \sum_{i=1}^n |x_i+y_i|^{p-1}|x_i| + \sum_{i=1}^n |x_i+y_i|^{p-1}|y_i| \\ &= \sum_{i=1}^n |x_i+y_i|^{\frac{p}{p'}}|x_i| + \sum_{i=1}^n |x_i+y_i|^{\frac{p}{p'}}|y_i| \\ &\leq \left(\sum_{i=1}^n |x_i+y_i|^p\right)^{1/p'}\left[\left(\sum_{i=1}^n |x_i|^p\right)^{1/p} + \left(\sum_{i=1}^n |y_i|^p\right)^{1/p}\right] \\ &= \|\boldsymbol{x}+\boldsymbol{y}\|^{p/p'}\left(\|\boldsymbol{x}\|_p + \|\boldsymbol{y}\|_p\right)\end{aligned}$$

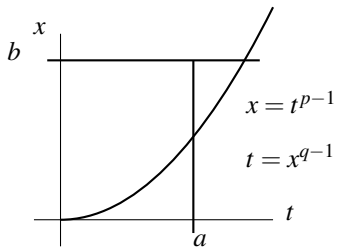
so dividing by $\|\boldsymbol{x}+\boldsymbol{y}\|^{p/p'}$, it follows

$$\|\boldsymbol{x}+\boldsymbol{y}\|^p\,\|\boldsymbol{x}+\boldsymbol{y}\|^{-p/p'} = \|\boldsymbol{x}+\boldsymbol{y}\| \leq \|\boldsymbol{x}\|_p + \|\boldsymbol{y}\|_p$$

$\left(p - \frac{p}{p'} = p\left(1-\frac{1}{p'}\right) = p\frac{1}{p} = 1.\right)$. ■

It only remains to prove Lemma 4.3.10.

Proof of the lemma: Let $p' = q$ to save on notation and consider the following picture:



$$ab \leq \int_0^a t^{p-1}dt + \int_0^b x^{q-1}dx = \frac{a^p}{p} + \frac{b^q}{q}.$$

Note equality occurs when $a^p = b^q$. ■

Alternate proof of the lemma: First note that if either a or b are zero, then there is nothing to show so we can assume $b, a > 0$. Let $b > 0$ and let

$$f(a) = \frac{a^p}{p} + \frac{b^q}{q} - ab$$

Then the second derivative of f is positive on $(0,\infty)$ so its graph is convex. Also $f(0) > 0$ and $\lim_{a\to\infty} f(a) = \infty$. Then a short computation shows that there is only one critical point, where f is minimized and this happens when a is such that $a^p = b^q$. At this point,

$$f(a) = b^q - b^{q/p}b = b^q - b^{q-1}b = 0$$

Therefore, $f(a) \geq 0$ for all a and this proves the lemma. ■

Another example of a very useful norm on $\mathbb{F}^n$ is the norm $\|\cdot\|_\infty$ defined by

$$\|\boldsymbol{x}\|_\infty \equiv \max\{|x_k| : k = 1,2,\cdots,n\}$$

You should verify that this satisfies all the axioms of a norm. Here is the triangle inequality.

$$\begin{aligned} \|\boldsymbol{x}+\boldsymbol{y}\|_\infty &= \max_k\{|x_k+y_k|\} \leq \max_k\{|x_k|+|y_k|\} \\ &\leq \max_k\{|x_k|\} + \max_k\{|y_k|\} = \|\boldsymbol{x}\|_\infty + \|\boldsymbol{y}\|_\infty \end{aligned}$$

It turns out that in terms of analysis, it makes **absolutely no difference** which norm you use. This will be explained later. First is a short review of the notion of orthonormal bases which is not needed directly in what follows but is sufficiently important to include.

4.3.5 Orthonormal Bases

Not all bases for an inner product space H are created equal. The best bases are orthonormal.

Definition 4.3.12 *Suppose $\{\boldsymbol{v}_1,\cdots,\boldsymbol{v}_k\}$ is a set of vectors in an inner product space H. It is an orthonormal set if*

$$(\boldsymbol{v}_i,\boldsymbol{v}_j) = \delta_{ij} = \begin{cases} 1 \text{ if } i = j \\ 0 \text{ if } i \neq j \end{cases}$$

Every orthonormal set of vectors is automatically linearly independent.

Proposition 4.3.13 *Suppose $\{\boldsymbol{v}_1,\cdots,\boldsymbol{v}_k\}$ is an orthonormal set of vectors. Then it is linearly independent.*

Proof: Suppose $\sum_{i=1}^k c_i\boldsymbol{v}_i = \boldsymbol{0}$. Then taking inner products with

$$\boldsymbol{v}_j, 0 = (\boldsymbol{0},\boldsymbol{v}_j) = \sum_i c_i(\boldsymbol{v}_i,\boldsymbol{v}_j) = \sum_i c_i\delta_{ij} = c_j.$$

Since j is arbitrary, this shows the set is linearly independent as claimed. ■

It turns out that if X is any subspace of H, then there exists an orthonormal basis for X. The process by which this is done is called the Gram Schmidt process.

Lemma 4.3.14 *Let X be a subspace of dimension n which is contained in an inner product space H. Let a basis for X be $\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\}$. Then there exists an orthonormal basis for X, $\{\boldsymbol{u}_1,\cdots,\boldsymbol{u}_n\}$ which has the property that for each $k\leq n$, span$(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_k)=$* $\operatorname{span}(\boldsymbol{u}_1,\cdots,\boldsymbol{u}_k)$.

Proof: Let $\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\}$ be a basis for X. Let $\boldsymbol{u}_1\equiv\boldsymbol{x}_1/|\boldsymbol{x}_1|$. Thus for $k=1$,

$$\operatorname{span}(\boldsymbol{u}_1)=\operatorname{span}(\boldsymbol{x}_1)$$

and $\{\boldsymbol{u}_1\}$ is an orthonormal set. Now suppose for some $k<n$, $\boldsymbol{u}_1,\cdots,\boldsymbol{u}_k$ have been chosen such that $(\boldsymbol{u}_j,\boldsymbol{u}_l)=\delta_{jl}$ and $\operatorname{span}(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_k)=\operatorname{span}(\boldsymbol{u}_1,\cdots,\boldsymbol{u}_k)$. Then define

$$\boldsymbol{u}_{k+1}\equiv\frac{\boldsymbol{x}_{k+1}-\sum_{j=1}^{k}(\boldsymbol{x}_{k+1},\boldsymbol{u}_j)\,\boldsymbol{u}_j}{\left|\boldsymbol{x}_{k+1}-\sum_{j=1}^{k}(\boldsymbol{x}_{k+1},\boldsymbol{u}_j)\,\boldsymbol{u}_j\right|},\tag{4.20}$$

where the denominator is not equal to zero because the $\boldsymbol{x}_j$ form a basis and so

$$\boldsymbol{x}_{k+1}\notin\operatorname{span}(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_k)=\operatorname{span}(\boldsymbol{u}_1,\cdots,\boldsymbol{u}_k)$$

Thus by induction,

$$\boldsymbol{u}_{k+1}\in\operatorname{span}(\boldsymbol{u}_1,\cdots,\boldsymbol{u}_k,\boldsymbol{x}_{k+1})=\operatorname{span}(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_k,\boldsymbol{x}_{k+1}).$$

Also, $\boldsymbol{x}_{k+1}\in\operatorname{span}(\boldsymbol{u}_1,\cdots,\boldsymbol{u}_k,\boldsymbol{u}_{k+1})$ which is seen easily by solving 4.20 for $\boldsymbol{x}_{k+1}$ and it follows

$$\operatorname{span}(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_k,\boldsymbol{x}_{k+1})=\operatorname{span}(\boldsymbol{u}_1,\cdots,\boldsymbol{u}_k,\boldsymbol{u}_{k+1}).$$

If $l\leq k$, then denoting by C the scalar $\left|\boldsymbol{x}_{k+1}-\sum_{j=1}^{k}(\boldsymbol{x}_{k+1},\boldsymbol{u}_j)\,\boldsymbol{u}_j\right|^{-1}$,

$$\begin{aligned}(\boldsymbol{u}_{k+1},\boldsymbol{u}_l) &= C\left((\boldsymbol{x}_{k+1},\boldsymbol{u}_l)-\sum_{j=1}^{k}(\boldsymbol{x}_{k+1},\boldsymbol{u}_j)(\boldsymbol{u}_j,\boldsymbol{u}_l)\right)\\ &= C\left((\boldsymbol{x}_{k+1},\boldsymbol{u}_l)-\sum_{j=1}^{k}(\boldsymbol{x}_{k+1},\boldsymbol{u}_j)\,\delta_{lj}\right)\\ &= C\left((\boldsymbol{x}_{k+1},\boldsymbol{u}_l)-(\boldsymbol{x}_{k+1},\boldsymbol{u}_l)\right)=0.\end{aligned}$$

The vectors, $\{\boldsymbol{u}_j\}_{j=1}^{n}$, generated in this way are therefore an orthonormal basis because each vector has unit length. ■

The process by which these vectors were generated is called the Gram Schmidt process.

4.4 Equivalence of Norms

As mentioned above, it makes absolutely no difference which norm you decide to use. This holds in general finite dimensional normed spaces. First are some simple lemmas featuring one dimensional considerations. In this case, the distance is given by $d(x,y)=|x-y|$ and so the open balls are sets of the form $(x-\delta,x+\delta)$.

Also recall the Lemma 3.5.9 which is stated next for convenience.

Lemma 4.4.1 *The closed interval $[a,b]$ is compact.*

Corollary 4.4.2 *The set $Q \equiv [a,b]+i[c,d] \subseteq \mathbb{C}$ is compact, meaning*

$$\{x+iy : x \in [a,b], y \in [c,d]\}$$

Proof: Let $\{x_n+iy_n\}$ be a sequence in Q. Then there is a subsequence such that $\lim_{k\to\infty} x_{n_k} = x \in [a,b]$. There is a further subsequence such that $\lim_{l\to\infty} y_{n_{k_l}} = y \in [c,d]$. Thus, also $\lim_{l\to\infty} x_{n_{k_l}} = x$ because subsequences of convergent sequences converge to the same point. Therefore, from the way we measure the distance in $\mathbb{C}$, it follows that $\lim_{l\to\infty}\left(x_{n_{k_l}}+y_{n_{k_l}}\right) = x+iy \in Q$. ■

The next corollary gives the definition of a closed disk and shows that, like a closed interval, a closed disk is compact.

Corollary 4.4.3 *In $\mathbb{C}$, let $D(z,r) \equiv \{w \in \mathbb{C} : |z-w| \le r\}$. Then $D(z,r)$ is compact.*

Proof: Note that

$$D(z,r) \subseteq [\mathrm{Re}\, z - r, \mathrm{Re}\, z + r] + i[\mathrm{Im}\, z - r, \mathrm{Im}\, z + r]$$

which was just shown to be compact. Also, if $w_k \to w$ where $w_k \in D(z,r)$, then by the triangle inequality,

$$|z-w| = \lim_{k\to\infty} |z-w_k| \le r$$

and so $D(z,r)$ is a closed subset of a compact set. Hence it is compact by Proposition 3.5.2. ■

Recall that sequentially compact and compact are the same in any metric space which is the context of the assertions here.

Lemma 4.4.4 *Let K_i be a nonempty compact set in $\mathbb{F}$. Then $P \equiv \prod_{i=1}^n K_i$ is compact in $\mathbb{F}^n$.*

Proof: Let $\{\boldsymbol{x}_k\}$ be a sequence in P. Taking a succession of subsequences as in the proof of Corollary 4.4.2, there exists a subsequence, still denoted as $\{\boldsymbol{x}_k\}$ such that if x_k^i is the i^{th} component of $\boldsymbol{x}_k$, then $\lim_{k\to\infty} x_k^i = x^i \in K_i$. Thus if $\boldsymbol{x}$ is the vector of P whose i^{th} component is x^i,

$$\lim_{k\to\infty} |\boldsymbol{x}_k - \boldsymbol{x}| \equiv \lim_{k\to\infty} \left(\sum_{i=1}^n \left|x_k^i - x^i\right|^2 \right)^{1/2} = 0$$

It follows that P is sequentially compact, hence compact. ■

A set K in $\mathbb{F}^n$ is said to be bounded if it is contained in some ball $B(\mathbf{0},r)$.

Theorem 4.4.5 *A set $K \subseteq \mathbb{F}^n$ is compact if it is closed and bounded. If $f : K \to \mathbb{R}$, then f achieves its maximum and its minimum on K.*

Proof: Say K is closed and bounded, being contained in $B(\mathbf{0},r)$. Then if $\boldsymbol{x} \in K$, $|x_i| < r$ where x_i is the i^{th} component. Hence $K \subseteq \prod_{i=1}^n D(0,r)$, a compact set by Lemma 4.4.4. By Proposition 3.5.2, since K is a closed subset of a compact set, it is compact. The last claim is just the extreme value theorem, Theorem 3.7.2. ■

Definition 4.4.6 *Let $\{v_1,\cdots,v_n\}$ be a basis for V where $(V,\|\cdot\|)$ is a finite dimensional normed vector space with field of scalars equal to either $\mathbb{R}$ or $\mathbb{C}$. Define $\theta : V \to \mathbb{F}^n$ as follows.*

$$\theta\left(\sum_{j=1}^{n}\alpha_j v_j\right) \equiv \alpha \equiv (\alpha_1,\cdots,\alpha_n)^T$$

Thus θ maps a vector to its coordinates taken with respect to a given basis.

The following fundamental lemma comes from the extreme value theorem for continuous functions defined on a compact set. Let

$$f(\alpha) \equiv \left\|\sum_i \alpha_i v_i\right\| \equiv \left\|\theta^{-1}\alpha\right\|$$

Then it is clear that f is a continuous function defined on $\mathbb{F}^n$. This is because $\alpha \to \sum_i \alpha_i v_i$ is a continuous map into V and from the triangle inequality $x \to \|x\|$ is continuous as a map from V to $\mathbb{R}$.

Lemma 4.4.7 *There exists $\delta > 0$ and $\Delta \geq \delta$ such that*

$$\delta = \min\{f(\alpha) : |\alpha| = 1\},\ \Delta = \max\{f(\alpha) : |\alpha| = 1\}$$

Also,

$$\delta|\alpha| \leq \left\|\theta^{-1}\alpha\right\| \leq \Delta|\alpha| \tag{4.21}$$

$$\delta|\theta v| \leq \|v\| \leq \Delta|\theta v| \tag{4.22}$$

Proof: These numbers exist thanks to Theorem 4.4.5. It cannot be that $\delta = 0$ because if it were, you would have $|\alpha| = 1$ but $\sum_{j=1}^n \alpha_k v_j = 0$ which is impossible since $\{v_1,\cdots,v_n\}$ is linearly independent. The first of the above inequalities follows from $\delta \leq \left\|\theta^{-1}\frac{\alpha}{|\alpha|}\right\| = f\left(\frac{\alpha}{|\alpha|}\right) \leq \Delta$. The second follows from observing that $\theta^{-1}\alpha$ is a generic vector v in V. ■

Note that these inequalities yield the fact that convergence of the coordinates with respect to a given basis is equivalent to convergence of the vectors. More precisely, to say that $\lim_{k\to\infty} v^k = v$ is the same as saying that $\lim_{k\to\infty} \theta v^k = \theta v$. Indeed,

$$\delta|\theta v_n - \theta v| \leq \|v_n - v\| \leq \Delta|\theta v_n - \theta v|$$

Now we can draw several conclusions about $(V,\|\cdot\|)$ for V finite dimensional.

Theorem 4.4.8 *Let $(V,\|\cdot\|)$ be a finite dimensional normed linear space. Then the compact sets are exactly those which are closed and bounded. Also $(V,\|\cdot\|)$ is complete. If K is a closed and bounded set in $(V,\|\cdot\|)$ and $f : K \to \mathbb{R}$, then f achieves its maximum and minimum on K.*

Proof: First note that the inequalities 4.21 and 4.22 show that both θ^{-1} and θ are continuous. Thus these take convergent sequences to convergent sequences.

Let $\{w_k\}_{k=1}^\infty$ be a Cauchy sequence. Then from 4.22, $\{\theta w_k\}_{k=1}^\infty$ is a Cauchy sequence. Thanks to Theorem 4.4.5, it converges to some $\beta \in \mathbb{F}^n$. It follows that $\lim_{k\to\infty}\theta^{-1}\theta w_k = \lim_{k\to\infty} w_k = \theta^{-1}\beta \in V$. This shows completeness.

Next let K be a closed and bounded set. Let $\{\boldsymbol{w}_k\} \subseteq K$. Then $\{\theta \boldsymbol{w}_k\} \subseteq \theta K$ which is also a closed and bounded set thanks to the inequalities 4.21 and 4.22. Thus there is a subsequence still denoted with k such that $\theta \boldsymbol{w}_k \to \boldsymbol{\beta} \in \mathbb{F}^n$. Then as just done, $\boldsymbol{w}_k \to \theta^{-1}\boldsymbol{\beta}$. Since K is closed, it follows that $\theta^{-1}\beta \in K$.

This has just shown that a closed and bounded set in V is sequentially compact hence compact.

Finally, why are the only compact sets those which are closed and bounded? Let K be compact. If it is not bounded, then there is a sequence of points of $K, \{\boldsymbol{k}^m\}_{m=1}^{\infty}$ such that $\|\boldsymbol{k}^m\| \geq \left\|\boldsymbol{k}^{m-1}\right\| + 1$. It follows that it cannot have a convergent subsequence because the points are further apart from each other than 1/2. Indeed,

$$\left\|\boldsymbol{k}^m - \boldsymbol{k}^{m+1}\right\| \geq \left\|\boldsymbol{k}^{m+1}\right\| - \|\boldsymbol{k}^m\| \geq 1 > 1/2$$

Hence K is not sequentially compact and consequently it is not compact. It follows that K is bounded. If K is not closed, then there exists a limit point $\boldsymbol{k}$ which is not in K. (Recall that closed means it has all its limit points.) By Theorem 3.1.8, there is a sequence of distinct points having no repeats and none equal to $\boldsymbol{k}$ denoted as $\{\boldsymbol{k}^m\}_{m=1}^{\infty}$ such that $\boldsymbol{k}^m \to \boldsymbol{k}$. Then this sequence $\{\boldsymbol{k}^m\}$ fails to have a subsequence which converges to a point of K. Hence K is not sequentially compact. Thus, if K is compact then it is closed and bounded.

The last part is the extreme value theorem, Theorem 3.7.2. ■

Next is the theorem which states that any two norms on a finite dimensional vector space are equivalent.

Theorem 4.4.9 *Let $\|\cdot\|, \|\cdot\|_1$ be two norms on V a finite dimensional vector space. Then they are equivalent, which means there are constants $0 < a < b$ such that for all $\boldsymbol{v}$,*

$$a\|\boldsymbol{v}\| \leq \|\boldsymbol{v}\|_1 \leq b\|\boldsymbol{v}\|$$

Proof: In Lemma 4.4.7, let δ, Δ go with $\|\cdot\|$ and $\hat{\delta}, \hat{\Delta}$ go with $\|\cdot\|_1$. Then using the inequalities of this lemma,

$$\|\boldsymbol{v}\| \leq \Delta|\theta \boldsymbol{v}| \leq \frac{\Delta}{\hat{\delta}}\|\boldsymbol{v}\|_1 \leq \frac{\Delta\hat{\Delta}}{\hat{\delta}}|\theta \boldsymbol{v}| \leq \frac{\Delta}{\delta}\frac{\hat{\Delta}}{\hat{\delta}}\|\boldsymbol{v}\|$$

and so $\frac{\hat{\delta}}{\Delta}\|\boldsymbol{v}\| \leq \|\boldsymbol{v}\|_1 \leq \frac{\hat{\Delta}}{\delta}\|\boldsymbol{v}\|$. Thus the norms are equivalent. ■

It follows right away that the closed and open sets are the same with two different norms. Also, all considerations involving limits are unchanged from one norm to another.

Corollary 4.4.10 *Consider the metric spaces $(V, \|\cdot\|_1), (V, \|\cdot\|_2)$ where V has dimension n. Then a set is closed or open in one of these if and only if it is respectively closed or open in the other. In other words, the two metric spaces have exactly the same open and closed sets. Also, a set is bounded in one metric space if and only if it is bounded in the other.*

Proof: This follows from Theorem 3.6.2, the theorem about the equivalent formulations of continuity. Using this theorem, it follows from Theorem 4.4.9 that the identity map $I(\boldsymbol{x}) \equiv \boldsymbol{x}$ is continuous. The reason for this is that the inequality of this theorem implies that if $\|\boldsymbol{v}^m - \boldsymbol{v}\|_1 \to 0$ then $\|I\boldsymbol{v}^m - I\boldsymbol{v}\|_2 = \|I(\boldsymbol{v}^m - \boldsymbol{v})\|_2 \to 0$ and the same holds on switching 1 and 2 in what was just written.

Therefore, the identity map takes open sets to open sets and closed sets to closed sets. In other words, the two metric spaces have the same open sets and the same closed sets.

Suppose S is bounded in $(V,\|\cdot\|_1)$. This means it is contained in $B(\mathbf{0},r)_1$ where the subscript of 1 indicates the norm is $\|\cdot\|_1$. Let $\delta\|\cdot\|_1 \leq \|\cdot\|_2 \leq \Delta\|\cdot\|_1$ as described above. Then $S \subseteq B(\mathbf{0},r)_1 \subseteq B(\mathbf{0},\Delta r)_2$ so S is also bounded in $(V,\|\cdot\|_2)$. Similarly, if S is bounded in $\|\cdot\|_2$ then it is bounded in $\|\cdot\|_1$. ■

One can show that in the case of $\mathbb{R}$ where it makes sense to consider sup and inf, convergence of Cauchy sequences can be shown to imply the other definition of completeness involving sup, and inf.

4.5 Covering Theorems

These covering theorems make sense on any finite dimensional normed linear space. There are two which are commonly used, the Vitali theorem and the Besicovitch theorem. The first adjusts the size of balls and the second does not. Of the two, it is the Besicovitch theorem which I will emphasize. However, the Vitali theorem is used more often and may be a little easier. I decided to place these theorems early in the book to emphasize that they only require a finite dimensional normed linear space.

4.5.1 Vitali Covering Theorem

The Vitali covering theorem is a profound result about coverings of a set in $(X,\|\cdot\|)$ with balls. Usually we are interested in $\mathbb{R}^p$ with some norm. We will tacitly assume all balls have positive radius. They will not be single points. Before beginning the proof, here is a useful lemma.

Lemma 4.5.1 *In a normed linear space, $\overline{B(\boldsymbol{x},r)} = \{\boldsymbol{y} : \|\boldsymbol{y}-\boldsymbol{x}\| \leq r\}$.*

Proof: It is clear that $\overline{B(\boldsymbol{x},r)} \subseteq \{\boldsymbol{y} : \|\boldsymbol{y}-\boldsymbol{x}\| \leq r\}$ because if $\boldsymbol{y} \in \overline{B(\boldsymbol{x},r)}$, then there exists a sequence of points of $B(\boldsymbol{x},r)$, $\{\boldsymbol{x}_n\}$ such that $\|\boldsymbol{x}_n-\boldsymbol{y}\| \to 0, \|\boldsymbol{x}_n\| < r$. However, this requires that $\|\boldsymbol{x}_n\| \to \|\boldsymbol{y}\|$ and so $\|\boldsymbol{y}\| \leq r$. Now let $\boldsymbol{y}$ be in the right side. It suffices to consider $\|\boldsymbol{y}-\boldsymbol{x}\| = 1$. Then you could consider for $t \in (0,1)$, $\boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x}) = \boldsymbol{z}(t)$. Then $\|\boldsymbol{z}(t)-\boldsymbol{x}\| = t\|\boldsymbol{y}-\boldsymbol{x}\| = tr < r$ and so $\boldsymbol{z}(t) \in B(\boldsymbol{x},r)$. But also, $\|\boldsymbol{z}(t)-\boldsymbol{y}\| = (1-t)\|\boldsymbol{y}-\boldsymbol{x}\| = (1-t)r$ so $\lim_{t\to 0}\|\boldsymbol{z}(t)-\boldsymbol{y}\| = 0$ showing that $\boldsymbol{y} \in \overline{B(\boldsymbol{x},r)}$. ■

Thus the usual way we think about the closure of a ball is completely correct in a normed linear space. Its limit points not in the ball are exactly $\boldsymbol{y}$ such that $\|\boldsymbol{y}-\boldsymbol{x}\| = r$. Recall that this lemma is not always true in the context of a metric space. Recall the discrete metric for example, in which the distance between different points is 1 and distance between a point and itself is 0. In what follows I will use the result of this lemma without comment. Balls will be either open, closed or neither. I am going to use the Hausdorff maximal theorem, Theorem 2.8.2 because it yields a very simple argument. It can be done other ways however. In the argument, the balls are not necessarily open nor closed. $\boldsymbol{y}$ is in $B(\boldsymbol{x},r)$ will mean that $\|\boldsymbol{y}-\boldsymbol{x}\| < r$ or $\|\boldsymbol{y}-\boldsymbol{x}\| = r$.

Lemma 4.5.2 *Let $\mathscr{F}$ be a nonempty collection of balls satisfying*

$$\infty > M \equiv \sup\{r : B(\boldsymbol{p},r) \in \mathscr{F}\} > 0$$

and let $k \in (0,M)$. Then there exists $\mathscr{G} \subseteq \mathscr{F}$ such that

$$\text{If } B(\boldsymbol{p},r) \in \mathscr{G}, \text{then } r > k, \tag{4.23}$$

$$\text{If } B_1, B_2 \in \mathscr{G} \text{ then } \overline{B_1} \cap \overline{B_2} = \emptyset, \tag{4.24}$$

$$\mathscr{G} \text{ is maximal with respect to 4.23 and 4.24.} \tag{4.25}$$

By this is meant that if $\mathscr{H}$ is a collection of balls satisfying 4.23 and 4.24, then $\mathscr{H}$ cannot properly contain $\mathscr{G}$.

Proof: Let $\mathfrak{S}$ denote a subset of $\mathscr{F}$ such that 4.23 and 4.24 are satisfied. Since $k < M$, 4.23 is satisfied for some ball of $\mathfrak{S}$. Thus $\mathfrak{S} \neq \emptyset$. Partially order $\mathfrak{S}$ with respect to set inclusion. Thus $\mathscr{A} \prec \mathscr{B}$ for $\mathscr{A}, \mathscr{B}$ in $\mathfrak{S}$ means that $\mathscr{A} \subseteq \mathscr{B}$. By the Hausdorff maximal theorem, there is a maximal chain in $\mathfrak{S}$ denoted by $\mathscr{C}$. Then let $\mathscr{G}$ be $\cup\mathscr{C}$. If B_1, B_2 are in $\mathscr{C}$, then since $\mathscr{C}$ is a chain, both B_1, B_2 are in some element of $\mathscr{C}$ and so $\overline{B_1} \cap \overline{B_2} = \emptyset$. The maximality of $\mathscr{C}$ is violated if there is any other element of $\mathfrak{S}$ which properly contains $\mathscr{G}$. ■

Proposition 4.5.3 *Let $\mathscr{F}$ be a collection of balls, and let*

$$A \equiv \cup\{B : B \in \mathscr{F}\}.$$

Suppose

$$\infty > M \equiv \sup\{r : B(\boldsymbol{p}, r) \in \mathscr{F}\} > 0.$$

Then there exists $\mathscr{G} \subseteq \mathscr{F}$ such that $\mathscr{G}$ consists of balls whose closures are disjoint and

$$A \subseteq \cup\{\widehat{B} : B \in \mathscr{G}\}$$

where for $B = B(\boldsymbol{x}, r)$ a ball, $\widehat{B}$ denotes the open ball $B(\boldsymbol{x}, 5r)$.

Proof: Let $\mathscr{G}_1$ satisfy 4.23 - 4.25 for $k = \frac{2M}{3}$.

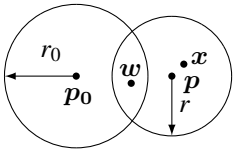
Suppose $\mathscr{G}_1, \cdots, \mathscr{G}_{m-1}$ have been chosen for $m \geq 2$. Let $\overline{\mathscr{G}_i}$ denote the collection of closures of the balls of $\mathscr{G}_i$. Then let $\mathscr{F}_m$ be those balls of $\mathscr{F}$, such that if B is one of these balls, $\overline{B}$ has empty intersection with every closed ball of $\overline{\mathscr{G}_i}$ for each $i \leq m-1$. Then using Lemma 4.5.2, let $\mathscr{G}_m$ be a maximal collection of balls from $\mathscr{F}_m$ with the property that each ball has radius larger than $\left(\frac{2}{3}\right)^m M$ and their closures are disjoint. Let $\mathscr{G} \equiv \cup_{k=1}^{\infty} \mathscr{G}_k$. Thus the closures of balls in $\mathscr{G}$ are disjoint. Let $\boldsymbol{x} \in B(\boldsymbol{p}, r) \in \mathscr{F} \setminus \mathscr{G}$. Choose m such that

$$\left(\frac{2}{3}\right)^m M < r \leq \left(\frac{2}{3}\right)^{m-1} M$$

Then $\overline{B(\boldsymbol{p}, r)}$ must have nonempty intersection with the closure of some ball from $\mathscr{G}_1 \cup \cdots \cup \mathscr{G}_m$ because if it didn't, then $\mathscr{G}_m$ would fail to be maximal. Denote by $B(\boldsymbol{p}_0, r_0)$ a ball in $\mathscr{G}_1 \cup \cdots \cup \mathscr{G}_m$ whose closure has nonempty intersection with $\overline{B(\boldsymbol{p}, r)}$. Thus both

$$r_0, r > \left(\frac{2}{3}\right)^m M, \text{ so } r \leq \left(\frac{2}{3}\right)^{m-1} M < \frac{3}{2} r_0$$

Consider the picture, in which $\boldsymbol{w} \in \overline{B(\boldsymbol{p}_0, r_0)} \cap \overline{B(\boldsymbol{p}, r)}$.



Then for $\boldsymbol{x} \in \overline{B(\boldsymbol{p},r)}$,

$$\|\boldsymbol{x}-\boldsymbol{p}_0\| \leq \|\boldsymbol{x}-\boldsymbol{p}\| + \|\boldsymbol{p}-\boldsymbol{w}\| + \overbrace{\|\boldsymbol{w}-\boldsymbol{p}_0\|}^{\leq r_0}$$

$$\leq r + r + r_0 \leq 2\overbrace{\left(\frac{2}{3}\right)^{m-1}}^{<\frac{3}{2}r_0} M + r_0 \leq 2\left(\frac{3}{2}r_0\right) + r_0 \leq 4r_0$$

Thus $B(\boldsymbol{p},r)$ is contained in $\overline{B(\boldsymbol{p}_0,4r_0)}$. It follows that the closures of the balls of $\mathscr{G}$ are disjoint and the set $\left\{\hat{B} : B \in \mathscr{G}\right\}$ covers A. ■

Note that this theorem does not depend on the underlying space being finite dimensional. However, it is typically used in this setting. The next theorem of Besicovitch depends in an essential way on X being finite dimensional because it exploits compactness and various constants originate explicitly from this compactness. However, no effort is being made here to give the most general conditions under which such covering theorems hold.

4.5.2 Besicovitch Covering Theorem

The covering theorems will have applications to measure theory presented later. In contrast to the Vitali covering theorem, one does not enlarge the balls in the Besicovitch covering theorem. This is extremely useful in the notion of general differentiation theorems for measures other than Lebesgue measure. The proof of this major result has to do with counting the number of times various balls can intersect. These estimates are used along with the pigeon hole principle to prove the result. This principle says that if you have n holes and $m > n$ pigeons, each of which must go in a hole, then some hole has more than one pigeon. In what follows $\boldsymbol{x}$ will continue to be in a normed linear space $(X, \|\cdot\|)$ of dimension p. This covering theorem is one of the most amazing and insightful ideas that I have ever encountered. It is simultaneously elegant, elementary and profound. This section is an attempt to present this wonderful result.

Here is a sequence of balls from $\mathscr{F}$ in the case that the set of centers of these balls is bounded. I will denote by $r(B_k)$ the radius of a ball B_k.

A construction of a sequence of balls

Lemma 4.5.4 *Let $\mathscr{F}$ be a nonempty set of nonempty balls in X with*

$$\sup\{\operatorname{diam}(B) : B \in \mathscr{F}\} = D < \infty$$

and let A denote the set of centers of these balls. Suppose A is bounded. Define a sequence of balls from $\mathscr{F}$, $\left\{B_j\right\}_{j=1}^{J}$ where $J \leq \infty$ such that

$$r(B_1) > \frac{3}{4}\sup\{r(B) : B \in \mathscr{F}\} \tag{4.26}$$

and if

$$A_m \equiv A \setminus (\cup_{i=1}^{m} B_i) \neq \emptyset, \tag{4.27}$$

then $B_{m+1} \in \mathscr{F}$ is chosen with center in A_m such that

$$r(B_m) > r(B_{m+1}) > \frac{3}{4}\sup\{r : B(\boldsymbol{a},r) \in \mathscr{F},\ \boldsymbol{a} \in A_m\}. \tag{4.28}$$

Then letting $B_j = B(\boldsymbol{a}_j, r_j)$, this sequence satisfies $\{B(\boldsymbol{a}_j, r_j/3)\}_{j=1}^{J}$ are disjoint,.

$$A \subseteq \cup_{i=1}^{J} B_i. \tag{4.29}$$

Proof: First note that B_{m+1} can be chosen as in 4.28. This is because the A_m are decreasing and so

$$\begin{aligned} & \frac{3}{4}\sup\{r : B(\boldsymbol{a},r) \in \mathscr{F},\ \boldsymbol{a} \in A_m\} \\ \leq\ & \frac{3}{4}\sup\{r : B(\boldsymbol{a},r) \in \mathscr{F},\ \boldsymbol{a} \in A_{m-1}\} < r(B_m) \end{aligned}$$

Thus the $r(B_k)$ are strictly decreasing and so no B_k contains a center of any other B_j.

If $\boldsymbol{x} \in B(\boldsymbol{a}_j, r_j/3) \cap B(\boldsymbol{a}_i, r_i/3)$ where these balls are two which are chosen by the above scheme such that $j > i$, then from what was just shown

$$\|\boldsymbol{a}_j - \boldsymbol{a}_i\| \leq \|\boldsymbol{a}_j - \boldsymbol{x}\| + \|\boldsymbol{x} - \boldsymbol{a}_i\| \leq \frac{r_j}{3} + \frac{r_i}{3} \leq \left(\frac{1}{3} + \frac{1}{3}\right) r_i = \frac{2}{3} r_i < r_i$$

and this contradicts the construction because $\boldsymbol{a}_j$ is not covered by $B(\boldsymbol{a}_i, r_i)$.

Finally consider the claim that $A \subseteq \cup_{i=1}^{J} B_i$. Pick B_1 satisfying 4.26. If

$$B_1, \cdots, B_m$$

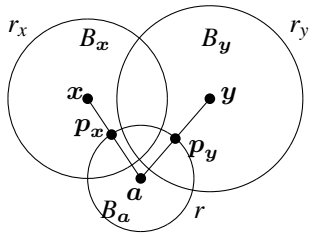
have been chosen, and A_m is given in 4.27, then if $A_m = \emptyset$, it follows $A \subseteq \cup_{i=1}^{m} B_i$. Set $J = m$.

Now let $\boldsymbol{a}$ be the center of $B_{\boldsymbol{a}} \in \mathscr{F}$. If $\boldsymbol{a} \in A_m$ for all m,(That is $\boldsymbol{a}$ does not get covered by the B_i.) then $r_{m+1} \geq \frac{3}{4} r(B_{\boldsymbol{a}})$ for all m, a contradiction since the balls $B\left(\boldsymbol{a}_j, \frac{r_j}{3}\right)$ are disjoint and A is bounded, implying that $r_j \to 0$. Thus $\boldsymbol{a}$ must fail to be in some A_m which means it was covered by some ball in the sequence. ■

The covering theorem is obtained by estimating how many B_j can intersect B_k for $j < k$. The thing to notice is that from the construction, no B_j contains the center of another B_i. Also, the $r(B_k)$ is a decreasing sequence.

Let $\alpha > 1$. There are two cases for an intersection. Either $r(B_j) \geq \alpha r(B_k)$ or $\alpha r(B_k) > r(B_j) > r(B_k)$.

First consider the case where we have a ball $B(\boldsymbol{a},r)$ intersected with other balls of radius larger than αr such that none of the balls contains the center of any other. This is illustrated in the following picture with two balls. This has to do with estimating the number of B_j for $j \leq k$ where $r(B_j) \geq \alpha r(B_k)$.



Imagine projecting the center of each big ball as in the above picture onto the surface of the given ball, assuming the given ball has radius 1. By scaling the balls, you could reduce to this case that the given ball has radius 1. Then from geometric reasoning, there should be a lower bound to the distance between these two projections depending on dimension. Thus there is an estimate on how many large balls can intersect the given ball with no ball containing a center of another one.

Intersections with relatively big balls

Lemma 4.5.5 *Let the balls $B_{\boldsymbol{a}}, B_{\boldsymbol{x}}, B_{\boldsymbol{y}}$ be as shown, having radii r, r_x, r_y respectively. Suppose the centers of $B_{\boldsymbol{x}}$ and $B_{\boldsymbol{y}}$ are not both in any of the balls shown, and suppose $r_y \geq r_x \geq \alpha r$ where α is a number larger than 1. Also let $P_{\boldsymbol{x}} \equiv \boldsymbol{a} + r\frac{\boldsymbol{x}-\boldsymbol{a}}{\|\boldsymbol{x}-\boldsymbol{a}\|}$ with $P_{\boldsymbol{y}}$ being defined similarly. Then it follows that $\|P_{\boldsymbol{x}} - P_{\boldsymbol{y}}\| \geq \frac{\alpha-1}{\alpha+1} r$. There exists a constant $L(p,\alpha)$ depending on α and the dimension, such that if $B_1, \cdots, B_m$ are all balls such that any pair are in the same situation relative to $B_{\boldsymbol{a}}$ as $B_{\boldsymbol{x}}$, and $B_{\boldsymbol{y}}$, then $m \leq L(p,\alpha)$.*

Proof: From the definition,

$$\begin{aligned}
\|P_{\boldsymbol{x}} - P_{\boldsymbol{y}}\| &= r\left\| \frac{\boldsymbol{x}-\boldsymbol{a}}{\|\boldsymbol{x}-\boldsymbol{a}\|} - \frac{\boldsymbol{y}-\boldsymbol{a}}{\|\boldsymbol{y}-\boldsymbol{a}\|} \right\| \\
&= r\left\| \frac{(\boldsymbol{x}-\boldsymbol{a})\|\boldsymbol{y}-\boldsymbol{a}\| - (\boldsymbol{y}-\boldsymbol{a})\|\boldsymbol{x}-\boldsymbol{a}\|}{\|\boldsymbol{x}-\boldsymbol{a}\|\,\|\boldsymbol{y}-\boldsymbol{a}\|} \right\| \\
&= r\left\| \frac{\|\boldsymbol{y}-\boldsymbol{a}\|(\boldsymbol{x}-\boldsymbol{y}) + (\boldsymbol{y}-\boldsymbol{a})(\|\boldsymbol{y}-\boldsymbol{a}\| - \|\boldsymbol{x}-\boldsymbol{a}\|)}{\|\boldsymbol{x}-\boldsymbol{a}\|\,\|\boldsymbol{y}-\boldsymbol{a}\|} \right\| \\
&\geq r\frac{\|\boldsymbol{x}-\boldsymbol{y}\|}{\|\boldsymbol{x}-\boldsymbol{a}\|} - r\frac{\|\boldsymbol{y}-\boldsymbol{a}\|\,|\|\boldsymbol{y}-\boldsymbol{a}\| - \|\boldsymbol{x}-\boldsymbol{a}\||}{\|\boldsymbol{x}-\boldsymbol{a}\|\,\|\boldsymbol{y}-\boldsymbol{a}\|} \\
&= r\frac{\|\boldsymbol{x}-\boldsymbol{y}\|}{\|\boldsymbol{x}-\boldsymbol{a}\|} - \frac{r}{\|\boldsymbol{x}-\boldsymbol{a}\|}\,|\|\boldsymbol{y}-\boldsymbol{a}\| - \|\boldsymbol{x}-\boldsymbol{a}\||\,. \qquad (4.30)
\end{aligned}$$

There are two cases. First suppose that $\|\boldsymbol{y}-\boldsymbol{a}\| - \|\boldsymbol{x}-\boldsymbol{a}\| \geq 0$. Then the above

$$= r\frac{\|\boldsymbol{x}-\boldsymbol{y}\|}{\|\boldsymbol{x}-\boldsymbol{a}\|} - \frac{r}{\|\boldsymbol{x}-\boldsymbol{a}\|}\,\|\boldsymbol{y}-\boldsymbol{a}\| + r.$$

From the assumptions, $\|\boldsymbol{x}-\boldsymbol{y}\| \geq r_y$ and also $\|\boldsymbol{y}-\boldsymbol{a}\| \leq r + r_y$. Hence the above

$$\begin{aligned}
&\geq r\frac{r_y}{\|\boldsymbol{x}-\boldsymbol{a}\|} - \frac{r}{\|\boldsymbol{x}-\boldsymbol{a}\|}(r+r_y) + r = r - r\frac{r}{\|\boldsymbol{x}-\boldsymbol{a}\|} \\
&\geq r\left(1 - \frac{r}{\|\boldsymbol{x}-\boldsymbol{a}\|}\right) \geq r\left(1 - \frac{r}{r_x}\right) \geq r\left(1 - \frac{1}{\alpha}\right) \geq r\frac{\alpha-1}{\alpha+1}.
\end{aligned}$$

The other case is that $\|\boldsymbol{y}-\boldsymbol{a}\| - \|\boldsymbol{x}-\boldsymbol{a}\| < 0$ in 4.30. Then in this case 4.30 equals

$$\begin{aligned}
&= r\left(\frac{\|\boldsymbol{x}-\boldsymbol{y}\|}{\|\boldsymbol{x}-\boldsymbol{a}\|} - \frac{1}{\|\boldsymbol{x}-\boldsymbol{a}\|}(\|\boldsymbol{x}-\boldsymbol{a}\| - \|\boldsymbol{y}-\boldsymbol{a}\|)\right) \\
&= \frac{r}{\|\boldsymbol{x}-\boldsymbol{a}\|}(\|\boldsymbol{x}-\boldsymbol{y}\| - (\|\boldsymbol{x}-\boldsymbol{a}\| - \|\boldsymbol{y}-\boldsymbol{a}\|))
\end{aligned}$$

Then since $\|\boldsymbol{x}-\boldsymbol{a}\| \leq r+r_x, \|\boldsymbol{x}-\boldsymbol{y}\| \geq r_y, \|\boldsymbol{y}-\boldsymbol{a}\| \geq r_y$, and remembering that $r_y \geq r_x \geq \alpha r$,

$$\begin{aligned} &\geq \frac{r}{r_x+r}\left(r_y-(r+r_x)+r_y\right) \geq \frac{r}{r_x+r}\left(r_y-(r+r_y)+r_y\right) \\ &\geq \frac{r}{r_x+r}\left(r_y-r\right) \geq \frac{r}{r_x+r}\left(r_x-r\right) \geq \frac{r}{r_x+\frac{1}{\alpha}r_x}\left(r_x-\frac{1}{\alpha}r_x\right) \\ &= \frac{r}{1+(1/\alpha)}\left(1-1/\alpha\right) = \frac{\alpha-1}{\alpha+1}r \end{aligned}$$

Replacing r with something larger, $\frac{1}{\alpha}r_x$ is justified by the observation that $x \to \frac{\alpha-x}{\alpha+x}$ is decreasing. This proves the estimate between $P_{\boldsymbol{x}}$ and $P_{\boldsymbol{y}}$.

Finally, in the case of the balls B_i having centers at $\boldsymbol{x}_i$, then as above, let $P_{\boldsymbol{x}_i} = \boldsymbol{a} + r\frac{\boldsymbol{x}_i-\boldsymbol{a}}{\|\boldsymbol{x}_i-\boldsymbol{a}\|}$. Then $\left(P_{\boldsymbol{x}_i}-\boldsymbol{a}\right)r^{-1}$ is on the unit sphere having center $\boldsymbol{0}$. Furthermore,

$$\left\|\left(P_{\boldsymbol{x}_i}-\boldsymbol{a}\right)r^{-1}-\left(P_{\boldsymbol{y}_i}-\boldsymbol{a}\right)r^{-1}\right\| = r^{-1}\left\|P_{\boldsymbol{x}_i}-P_{\boldsymbol{y}_i}\right\| \geq r^{-1}r\frac{\alpha-1}{\alpha+1} = \frac{\alpha-1}{\alpha+1}.$$

How many points on the unit sphere can be pairwise this far apart? The unit sphere is compact and so there exists a $\frac{1}{4}\left(\frac{\alpha-1}{\alpha+1}\right)$ net having $L(p,\alpha)$ points. Thus m cannot be any larger than $L(p,\alpha)$ because if it were, then by the pigeon hole principal, two of the points $\left(P_{\boldsymbol{x}_i}-\boldsymbol{a}\right)r^{-1}$ would lie in a single ball $B\left(p,\frac{1}{4}\left(\frac{\alpha-1}{\alpha+1}\right)\right)$ so they could not be $\frac{\alpha-1}{\alpha+1}$ apart. ■

The above lemma has to do with balls which are relatively large intersecting a given ball. Next is a lemma which has to do with relatively small balls intersecting a given ball. First is another lemma.

Lemma 4.5.6 *Let $\Gamma > 1$ and $B(\boldsymbol{a},\Gamma r)$ be a ball and suppose $\{B(\boldsymbol{x}_i,r_i)\}_{i=1}^m$ are balls contained in $B(\boldsymbol{a},\Gamma r)$ such that $r \leq r_i$ and none of these balls contains the center of another ball. Then there is a constant $M(p,\Gamma)$ such that $m \leq M(p,\Gamma)$.*

Proof: Let $\boldsymbol{z}_i = \boldsymbol{x}_i - \boldsymbol{a}$. Then $B(\boldsymbol{z}_i,r_i)$ are balls contained in $B(\boldsymbol{0},\Gamma r)$ with no ball containing a center of another. Then $B\left(\frac{\boldsymbol{z}_i}{\Gamma r},\frac{r_i}{\Gamma r}\right)$ are balls in $B(\boldsymbol{0},1)$ with no ball containing the center of another. By compactness, there is a $\frac{1}{8\Gamma}$ net for $\overline{B(\boldsymbol{0},1)}$, $\{\boldsymbol{y}_i\}_{i=1}^{M(p,\Gamma)}$. Thus the balls $B\left(\boldsymbol{y}_i,\frac{1}{8\Gamma}\right)$ cover $\overline{B(\boldsymbol{0},1)}$. If $m \geq M(p,\Gamma)$, then by the pigeon hole principle, one of these $B\left(\boldsymbol{y}_i,\frac{1}{8\Gamma}\right)$ would contain some $\frac{\boldsymbol{z}_i}{\Gamma r}$ and $\frac{\boldsymbol{z}_j}{\Gamma r}$ which requires $\left\|\frac{\boldsymbol{z}_i}{\Gamma r}-\frac{\boldsymbol{z}_j}{\Gamma r}\right\| \leq \frac{1}{4\Gamma} < \frac{r_j}{4\Gamma r}$ so $\frac{\boldsymbol{z}_i}{\Gamma r} \in B\left(\frac{\boldsymbol{z}_j}{\Gamma r},\frac{r_j}{\Gamma r}\right)$. Thus $m \leq M(p,\gamma,\Gamma)$. ■

Intersections with small balls

Lemma 4.5.7 *Let B be a ball having radius r and suppose B has nonempty intersection with the balls $B_1,\cdots,B_m$ having radii $r_1,\cdots,r_m$ respectively, and as before, no B_i contains the center of any other and the centers of the B_i are not contained in B. Suppose $\alpha > 1$ and $r \leq \min(r_1,\cdots,r_m)$, each $r_i < \alpha r$. Then there exists a constant $M(p,\alpha)$ such that $m \leq M(p,\alpha)$.*

Proof: Let $B = B(\boldsymbol{a},r)$. Then each B_i is contained in $B(\boldsymbol{a},2r+\alpha r+\alpha r)$. This is because if $\boldsymbol{y} \in B_i \equiv B(\boldsymbol{x}_i,r_i)$,

$$\|\boldsymbol{y}-\boldsymbol{a}\| \leq \|\boldsymbol{y}-\boldsymbol{x}_i\| + \|\boldsymbol{x}_i-\boldsymbol{a}\| \leq r_i + r + r_i < 2r+\alpha r+\alpha r$$

Thus B_i does not contain the center of any other B_j, these balls are each contained in $B(\mathbf{a}, r(2\alpha+2))$, and each radius is at least as large as r. By Lemma 4.5.6 there is a constant $M(p,\alpha)$ such that $m \leq M(p,\alpha)$. ■

Now here is the Besicovitch covering theorem. In the proof, we are considering the sequence of balls described above.

Theorem 4.5.8 *There exists a constant N_p, depending only on p with the following property. If $\mathscr{F}$ is any collection of nonempty balls in X with*

$$\sup\{\operatorname{diam}(B) : B \in \mathscr{F}\} < D < \infty$$

and if A is the set of centers of the balls in $\mathscr{F}$, then there exist subsets of $\mathscr{F}$, $\mathscr{H}_1, \cdots, \mathscr{H}_{N_p}$, such that each $\mathscr{H}_i$ is a countable collection of disjoint balls from $\mathscr{F}$ (possibly empty) and

$$A \subseteq \cup_{i=1}^{N_p} \cup \{B : B \in \mathscr{H}_i\}.$$

Proof: To begin with, suppose A is bounded. Let $L(p,\alpha)$ be the constant of Lemma 4.5.5 and let $M_p = L(p,\alpha) + M(p,\alpha) + 1$. Define the following sequence of subsets of $\mathscr{F}$, $\mathscr{G}_1, \mathscr{G}_2, \cdots, \mathscr{G}_{M_p}$. Referring to the sequence $\{B_k\}$ considered in Lemma 4.5.4, let $B_1 \in \mathscr{G}_1$ and if $B_1, \cdots, B_m$ have been assigned, each to a $\mathscr{G}_i$, place B_{m+1} in the first $\mathscr{G}_j$ such that B_{m+1} intersects no set already in $\mathscr{G}_j$. The existence of such a j follows from Lemmas 4.5.5 and 4.5.7 and the pigeon hole principle. Here is why. B_{m+1} can intersect at most $L(p,\alpha)$ sets of $\{B_1, \cdots, B_m\}$ which have radii at least as large as $\alpha r(B_{m+1})$ thanks to Lemma 4.5.5. It can intersect at most $M(p,\alpha)$ sets of $\{B_1, \cdots, B_m\}$ which have radius smaller than $\alpha r(B_{m+1})$ thanks to Lemma 4.5.7. Thus each $\mathscr{G}_j$ consists of disjoint sets of $\mathscr{F}$ and the set of centers is covered by the union of these $\mathscr{G}_j$. This proves the theorem in case the set of centers is bounded.

Now let $R_1 = B(\mathbf{0}, 5D)$ and if R_m has been chosen, let

$$R_{m+1} = B(\mathbf{0}, (m+1)5D) \setminus R_m$$

Thus, if $|k-m| \geq 2$, no ball from $\mathscr{F}$ having nonempty intersection with R_m can intersect any ball from $\mathscr{F}$ which has nonempty intersection with R_k. This is because all these balls have radius less than D. Now let $A_m \equiv A \cap R_m$ and apply the above result for a bounded set of centers to those balls of $\mathscr{F}$ which intersect R_m to obtain sets of disjoint balls $\mathscr{G}_1(R_m), \mathscr{G}_2(R_m), \cdots, \mathscr{G}_{M_p}(R_m)$ covering A_m. Then simply define $\mathscr{G}'_j \equiv \cup_{k=1}^{\infty} \mathscr{G}_j(R_{2k}), \mathscr{G}_j \equiv \cup_{k=1}^{\infty} \mathscr{G}_j(R_{2k-1})$. Let $N_p = 2M_p$ and

$$\left\{\mathscr{H}_1, \cdots, \mathscr{H}_{N_p}\right\} \equiv \left\{\mathscr{G}'_1, \cdots, \mathscr{G}'_{M_p}, \mathscr{G}_1, \cdots, \mathscr{G}_{M_p}\right\}$$

Note that the balls in $\mathscr{G}'_j$ are disjoint. This is because those in $\mathscr{G}_j(R_{2k})$ are disjoint and if you consider any ball in $\mathscr{G}_j(R_{2m})$, it cannot intersect a ball of $\mathscr{G}_j(R_{2k})$ for $m \neq k$ because $|2k-2m| \geq 2$. Similar considerations apply to the balls of $\mathscr{G}_j$. ■

Of course, you could pick a particular α. If you make α larger, $L(p,\alpha)$ should get smaller and $M(p,\alpha)$ should get larger. Obviously one could explore this at length to try and get a best choice of α.

4.6 Exercises

1. Let V be a vector space with basis $\{v_1, \cdots, v_n\}$. For $v \in V$, denote its coordinate vector as $\boldsymbol{v} = (\alpha_1, \cdots, \alpha_n)$ where $v = \sum_{k=1}^n \alpha_k v_k$. Now define

$$\|v\| \equiv \max\{|\alpha_k| : k = 1, ..., n\}.$$

   Show that this is a norm on V.

2. Let $(X, \|\cdot\|)$ be a normed linear space. You can let it be $(\mathbb{R}^n, |\cdot|)$ if you like. Recall $|\boldsymbol{x}|$ is the usual magnitude of a vector given by $|\boldsymbol{x}| = \sqrt{\sum_{k=1}^n |x_k|^2}$. A set A is said to be **convex** if whenever $\boldsymbol{x}, \boldsymbol{y} \in A$ the line segment determined by these points given by $t\boldsymbol{x} + (1-t)\boldsymbol{y}$ for $t \in [0,1]$ is also in A. Show that every open or closed ball is convex. Remember a closed ball is $D(\boldsymbol{x}, r) \equiv \{\hat{\boldsymbol{x}} : \|\hat{\boldsymbol{x}} - \boldsymbol{x}\| \leq r\}$ while the open ball is $B(\boldsymbol{x}, r) \equiv \{\hat{\boldsymbol{x}} : \|\hat{\boldsymbol{x}} - \boldsymbol{x}\| < r\}$. This should work just as easily in any normed linear space with any norm.

3. This problem is for those who have had a course in Linear algebra. A vector $\boldsymbol{v}$ is in the convex hull of S if there are finitely many vectors of $S, \{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_m\}$ and nonnegative scalars $\{t_1, \cdots, t_m\}$ such that $\boldsymbol{v} = \sum_{k=1}^m t_k \boldsymbol{v}_k$, $\sum_{k=1}^m t_k = 1$.Such a linear combination is called a convex combination. Suppose now that $S \subseteq V$, a vector space of dimension n. Show that if $\boldsymbol{v} = \sum_{k=1}^m t_k \boldsymbol{v}_k$ is a vector in the convex hull for $m > n+1$, then there exist other nonnegative scalars $\{t_k'\}$ summing to 1 such that $\boldsymbol{v} = \sum_{k=1}^{m-1} t_k' \boldsymbol{v}_k$.Thus every vector in the convex hull of S can be obtained as a convex combination of at most $n+1$ points of S. This incredible result is in Rudin [50]. Convexity is more a geometric property than a topological property. **Hint:** Consider $L : \mathbb{R}^m \to V \times \mathbb{R}$ defined by $L(\boldsymbol{a}) \equiv (\sum_{k=1}^m a_k \boldsymbol{v}_k, \sum_{k=1}^m a_k)$ Explain why $\ker(L) \neq \{\boldsymbol{0}\}$. This will involve observing that $\mathbb{R}^m$ has higher dimension that $V \times \mathbb{R}$. Thus L cannot be one to one because one to one functions take linearly independent sets to linearly independent sets and you can't have a linearly independent set with more than $n+1$ vectors in $V \times \mathbb{R}$. Next, letting $\boldsymbol{a} \in \ker(L) \setminus \{\boldsymbol{0}\}$ and $\lambda \in \mathbb{R}$, note that $\lambda\, \boldsymbol{a} \in \ker(L)$. Thus for all $\lambda \in \mathbb{R}$, $\boldsymbol{v} = \sum_{k=1}^m (t_k + \lambda a_k) \boldsymbol{v}_k$. Now vary λ till some $t_k + \lambda a_k = 0$ for some $a_k \neq 0$. You can assume each $t_k > 0$ since otherwise, there is nothing to show. This is a really nice result because it can be used to show that the convex hull of a compact set is also compact. You might try to show this if you feel like it.

4. Show that the usual norm in $\mathbb{F}^n$ given by $|\boldsymbol{x}| = (\boldsymbol{x}, \boldsymbol{x})^{1/2}$ satisfies the following identities, the first of them being the parallelogram identity and the second being the polarization identity.

$$\begin{aligned} |\boldsymbol{x}+\boldsymbol{y}|^2 + |\boldsymbol{x}-\boldsymbol{y}|^2 &= 2|\boldsymbol{x}|^2 + 2|\boldsymbol{y}|^2 \\ \mathrm{Re}(\boldsymbol{x}, \boldsymbol{y}) &= \frac{1}{4}\left(|\boldsymbol{x}+\boldsymbol{y}|^2 - |\boldsymbol{x}-\boldsymbol{y}|^2\right) \end{aligned}$$

   Show that these identities hold in any inner product space, not just $\mathbb{F}^n$.

5. Suppose K is a compact subset of (X, d) a metric space. Also let $\mathscr{C}$ be an open cover of K. Show that there exists $\delta > 0$ such that for all $x \in K$, $B(x, \delta)$ is contained in a single set of $\mathscr{C}$. This number is called a Lebesgue number. **Hint:** For each $x \in K$, there exists $B(x, \delta_x)$ such that this ball is contained in a set of $\mathscr{C}$. Now consider

the balls $\left\{B\left(x,\frac{\delta_x}{2}\right)\right\}_{x\in K}$. Finitely many of these cover K. $\left\{B\left(x_i,\frac{\delta_{x_i}}{2}\right)\right\}_{i=1}^{n}$ Now consider what happens if you let $\delta \leq \min\left\{\frac{\delta_{x_i}}{2}, i=1,2,\cdots,n\right\}$. Explain why this works. You might draw a picture to help get the idea.

6. Suppose $\mathscr{C}$ is a set of compact sets in a metric space (X,d) and suppose that the intersection of **every** finite subset of $\mathscr{C}$ is nonempty. This is called the **finite intersection property.** Show that $\cap\mathscr{C}$, the intersection of all sets of $\mathscr{C}$ is nonempty. This particular result is enormously important. **Hint:** You could let $\mathscr{U}$ denote the set $\left\{K^C : K \in \mathscr{C}\right\}$. If $\cap\mathscr{C}$ is empty, then its complement is $\cup\mathscr{U} = X$. Picking $K \in \mathscr{C}$, it follows that $\mathscr{U}$ is an open cover of K. $K \subseteq \cup_{i=1}^{m} K_i^C = \left(\cap_{i=1}^{m} K_i\right)^C$ Therefore, you would need to have $\left\{K_1^C,\cdots,K_m^C\right\}$ is a cover of K. In other words, Now what does this say about the intersection of K with these K_i?

7. If (X,d) is a compact metric space and $f : X \to Y$ is continuous where (Y,ρ) is another metric space, show that if f is continuous on X, then it is uniformly continuous. Recall that this means that if $\varepsilon > 0$ is given, then there exists $\delta > 0$ such that if $d(x,\hat{x}) < \delta$, then $\rho(f(x),f(\hat{x})) < \varepsilon$. Compare with the definition of continuity. **Hint:** If this is not so, then there exists $\varepsilon > 0$ and $x_n, \hat{x}_n$ such that $d(x_n,\hat{x}_n) < 1/n$ but $\rho(f(x_n),f(\hat{x}_n)) \geq \varepsilon$. Now use compactness to get a contradiction.

8. Prove the above problem using another approach. Use the existence of the Lebesgue number in Problem 5 to prove continuity on a compact set K implies uniform continuity on this set. **Hint:** Consider $\mathscr{C} \equiv \left\{f^{-1}(B(f(x),\varepsilon/2)) : x \in X\right\}$. This is an open cover of X. Let δ be a Lebesgue number for this open cover. Suppose $d(x,\hat{x}) < \delta$. Then both $x,\hat{x}$ are in $B(x,\delta)$ and so both are in $f^{-1}\left(B\left(f(\bar{x}),\frac{\varepsilon}{2}\right)\right)$. Hence

$$\rho(f(x),f(\bar{x})) < \frac{\varepsilon}{2}, \rho(f(\hat{x}),f(\bar{x})) < \frac{\varepsilon}{2}.$$

Now consider the triangle inequality.

9. Let X be a vector space. A Hamel basis is a subset of X, Λ such that every vector of X can be written as a finite linear combination of vectors of Λ and the vectors of Λ are linearly independent in the sense that if $\{x_1,\cdots,x_n\} \subseteq \Lambda$ and $\sum_{k=1}^{n} c_k x_k = 0$ then each $c_k = 0$. Using the Hausdorff maximal theorem, show that every non-zero vector space has a Hamel basis. **Hint:** Let $x_1 \neq 0$. Let $\mathscr{F}$ denote the collection of subsets of X, Λ containing x_1 with the property that the vectors of Λ are linearly independent. Partially order $\mathscr{F}$ by set inclusion and consider the union of a maximal chain.

10. Suppose X is a nonzero real or complex normed linear space and let

$$V = \text{span}(w_1,...,w_m)$$

where $\{w_1,...,w_m\}$ is a linearly independent set of vectors of X. Show that V is a closed subspace of X with $V \subsetneq X$. First explain why Theorem 4.2.11 implies any finite dimensional subspace of X can be written this way. **Hint:** You might want to use something like Lemma 4.4.7 to show this.

11. Suppose X is a normed linear space and its dimension is either infinite or greater than m where $V \equiv \text{span}(w_1,...,w_m)$ for $\{w_1,...,w_m\}$ an independent set of vectors of X.

Show $X \setminus V$ is a dense open subset of X which is equivalent to V containing no ball $B(v,r), \{w : \|w-v\| < r\}$. **Hint:** If $B(x,r)$ is contained in V, then show, that since V is a subspace, $B(0,r)$ is contained in V. Then show this implies $X \subseteq V$ which is not the case.

12. Show that if (X,d) is a metric space and H,K are disjoint closed sets, there are open sets U_H, U_K such that $H \subseteq U_H, K \subseteq U_K$ and $U_H \cap U_K = \emptyset$. **Hint:** Let $k \in K$. Explain why $\text{dist}(k,H) \equiv \inf\{\|k-h\| : h \in H\} \equiv 2\delta_k > 0$. Now consider $U_K \equiv \cup_{k\in K} B(k,\delta_k)$. Do something similar for $h \in H$ and consider $U_H \equiv \cup_{k\in H} B(h,\delta_h)$.

13. If, in a metric space, $B(p,\delta)$ is a ball, show that

$$\overline{B(p,\delta)} \subseteq D(p,\delta) \equiv \{x : \|x-p\| \leq \delta\}$$

Now suppose (X,d) is a complete metric space and $U_n, n \in \mathbb{N}$ is a dense open set in X. Also let W be any nonempty open set. Show there exists a ball $B_1 \equiv B(p_1,r_1)$ having radius smaller than 2^{-1} such that $\overline{B_1} \subseteq U_1 \cap W_1$. Next show there exists $B_2 \equiv B(p_2,r_2)$ such that $\overline{B_2} \subseteq B_1 \cap U_2 \cap W$ with the radius of B_2 less than 2^{-2}. Continue this way. Explain why $\{p_n\}_{n=1}^{\infty}$ is a Cauchy sequence converging to some $p \in W \cap (\cup_{n=1}^{\infty} U_n)$. This is the very important Baire theorem which says that in a complete metric space, the intersection of dense open sets is dense.

14. Suppose you have a complete normed linear space, $(X, \|\cdot\|)$. Use the above problems leading to the Baire theorem in 13 to show that if $\mathscr{B}$ is a Hamel basis for for X, then $\mathscr{B}$ cannot be countable. **Hint:** If $\mathscr{B} = \{v_i\}_{i=1}^{\infty}$, consider $V_n \equiv \text{span}(v_1,...,v_n)$. Then use a problem listed above to argue that V_n^C is a dense open set. Now apply Problem 13. This shows why the idea of a Hamel basis often fails to be very useful whereas, in finite dimensional settings, it is just what is needed.

15. In any complete normed linear space which is infinite dimensional, show the unit ball is not compact. Do this by showing the existence of a sequence which cannot have a convergent subsequence. **Hint:** Pick $\|x_1\| = 1$. Suppose $x_1,...,x_n$ have been chosen, each $\|x_k\| = 1$. Then there is $x \notin \text{span}(x_1,...,x_n) \equiv V_n$. Now consider v such that $\|x-v\| \leq \frac{3}{2}\text{dist}(x,V_n)$. Then argue that for $k \leq n$,

$$\left\| \frac{x-v}{\|x-v\|} - x_k \right\| = \left\| \frac{x - \left(\overbrace{v + \|x-v\| x_k}^{\in V_n} \right)}{\|x-v\|} \right\| \geq \frac{\text{dist}(x,V_n)}{(3/2)\,\text{dist}(x,V_n)} = \frac{2}{3}$$

16. Let X be a complete inner product space. Let $\mathscr{F}$ denote subsets $\beta \subseteq X$ such that whenever $x,y \in X, (x,y) = 0$ if $x \neq y$ and $(x,x) = 1$ if $x = y$. Thus these β are orthonormal sets. Show there exists a maximal orthonormal set. If X is separable, show that this maximal orthonormal set is countable. **Hint:** Use the Hausdorff maximal theorem. The next few problems involve linear algebra.

17. Let X be a real inner product space and let $\{v_1,...,v_n\}$ be vectors in X. Let G be the $n \times n$ matrix $G_{ij} \equiv (v_i, v_j)$. Show that G^{-1} exists if and only if $\{v_1,...,v_n\}$ is linearly independent. G is called the Grammian or the metric tensor.

18. ↑Let X be as above, a real inner product space, and let $V \equiv \text{span}(v_1,...,v_n)$. Let $u \in X$ and $z \in V$. Show that $|u-z| = \inf\{|u-v| : v \in V\}$ if and only if $(u-z,v_i) = 0$ for all v_i. Note that the v_i might not be linearly independent. Also show that $|u-z|^2 = |u|^2 - (z,u)$.

19. ↑ Let G be the matrix of Problem 17 where $\{v_1,...,v_n\}$ is linearly independent and $V \equiv \text{span}(v_1,...,v_n) \subseteq X$, an inner product space. Let $x \equiv \sum_i x^i v_i, y \equiv \sum_i y^i v_i$ be two vectors of V. Show that $(x,y) = \sum_{i,j} x^i G_{ij} x^j$. Show that $z \equiv \sum_i z^i v_i, z$ is closest to $u \in X$ if and only if for all $i = 1,...,n, (u,v_i) = \sum_j G_{ij} z^j$. This gives a system of linear equations which must be satisfied by the z^i in order that z just given is the best approximation to u. Next show that there exists such a solution thanks to Problem 17 which says that the matrix G is invertible, and if G^{-1} has ij^{th} component G^{ij}, one finds that $\sum_j G^{ij}(u,v_j) = z^i$.

20. ↑ In the situation of the above problems, suppose A is an $m \times n$ matrix. Use Problem 18 to show that for $\boldsymbol{y} \in \mathbb{R}^m$, there always exists a solution $\boldsymbol{x}$ to the system of equations $A^T\boldsymbol{y} = A^TA\boldsymbol{x}$. Explain how this is in a sense the best you can do to solve $\boldsymbol{y} = A\boldsymbol{x}$ even though this last system of equations might not have a solution. Here A^T is the transpose of the matrix A. The equations $A^T\boldsymbol{y} = A^TA\boldsymbol{x}$ are called the normal equations for the least squares problem. **Hint:** Verify that $\left(A^T\boldsymbol{y},\boldsymbol{x}\right) = \left(\boldsymbol{y},A\boldsymbol{x}\right)$. Let the subspace V be $A(\mathbb{R}^n)$, the vectors spanning it being $\{Ae_1,...,Ae_n\}$. From the above problem, there exists $A\boldsymbol{x}$ in V which is closest to $\boldsymbol{y}$. Now use the characterization of this vector $\left(\boldsymbol{y}-A\boldsymbol{x},A\boldsymbol{z}\right) = 0$ for all $\boldsymbol{z} \in \mathbb{R}^n, A\boldsymbol{z}$ being a generic vector in $A(\mathbb{R}^n)$.

21. ↑As an example of an inner product space, consider $C([0,1])$ with the inner product $\int_0^1 f(x)\,g(x)\,dx$ where this is the ordinary integral from calculus. Abusing notation, let $\{x^{p_1},...,x^{p_n}\}$ with $-\frac{1}{2} < p_1 < \cdots < p_n$ be functions, (vectors) in $C([0,1])$. Verify that these vectors are linearly independent. **Hint:** You might want to use the Cauchy identity in the appendix, Theorem 1.9.28.

22. ↑As above, if $\{v_1,...,v_n\}$ is linearly independent, the Grammian is $G = G(v_1,...,v_n)$, $G_{ij} \equiv (v_i,v_j)$, then if $u \notin \text{span}(v_1,...,v_n) \equiv V$ you could consider $G(v_1,...,v_n,u)$. Then if $d \equiv \min\{|u-v| : v \in \text{span}(v_1,...,v_n)\}$, show that $d^2 = \frac{\det G(v_1,...,v_n,u)}{\det G(v_1,...,v_n)}$. Justify the following steps. Letting z be the closest point of V to u, from the above, $\left(u-\sum_{i=1}^n z^i v_i, v_p\right) = 0$ for each v_p and so

$$(u,v_p) = \sum_{i=1}^n (v_p,v_i)\,z^i \qquad (*)$$

Also, since $(u-z,v) = 0$ for all $v \in V, |u|^2 = |u-z+z|^2 = |u-z|^2 + |z|^2$ so

$$|u|^2 = \left|u - \sum_{i=1}^n z^i v_i\right|^2 + \left|\sum_{i=1}^n z^i v_i\right|^2 = d^2 + \left|\sum_{i=1}^n z^i v_i\right|^2$$

$$= d^2 + \sum_j \overbrace{\sum_i (v_j,v_i)\,z^i}^{=(u,v_j)} z^j = d^2 + \sum_j (u,v_j)\,z^j$$

$$= d^2 + \boldsymbol{y}^T \boldsymbol{z},\ \boldsymbol{y} \equiv ((u,v_1),\cdots,(u,v_n))^T,\ \boldsymbol{z} \equiv \left(z^1,\cdots,z^n\right)^T$$

From $*$, $G\boldsymbol{z} = \boldsymbol{y}$, $\begin{pmatrix} G(v_1,...,v_n) & \boldsymbol{0} \\ \boldsymbol{y}^T & 1 \end{pmatrix} \begin{pmatrix} \boldsymbol{z} \\ d^2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{y} \\ \|u\|^2 \end{pmatrix}$. Now use Cramer's rule to solve for d^2 and get

$$d^2 = \frac{\det \begin{pmatrix} G(v_1,...,v_n) & \boldsymbol{y} \\ \boldsymbol{y}^T & |u|^2 \end{pmatrix}}{\det(G(v_1,...,v_n))} \equiv \frac{\det G(v_1,...,v_n,u)}{\det G(v_1,...,v_n)}$$

23. In the situation of Problem 21, let $f_k(x) \equiv x^k$ and let $V \equiv \operatorname{span}(f_{p_1},...,f_{p_n})$. give an estimate for the distance d between f_m and V for m a nonnegative integer and as in the above problem $-\frac{1}{2} < p_1 < \cdots < p_n$. Use Theorem 1.9.28 in the appendix and the above problem with $v_i \equiv f_{p_i}$ and $v_{n+1} \equiv f_m$. Justify the following manipulations. The numerator in the above formula for the distance is of the form $\frac{\prod_{j<i\le n+1}(p_i-p_j)^2}{\prod_{i,j\le n+1}(p_i+p_j+1)}$

$$= \frac{\prod_{j<i\le n}(p_i-p_j)^2 \prod_{j\le n}(m-p_j)^2}{\prod_{i,j\le n}(p_i+p_j+1)\prod_{i=1}^n(p_i+m+1)\prod_{j=1}^n(p_j+m+1)(2m+1)}$$

While $G(f_{p_1},...,f_{p_n}) = \frac{\prod_{j<i\le n}(p_i-p_j)^2}{\prod_{i,j\le n}(p_i+p_j+1)}$. Thus $d = \frac{\prod_{j\le n}|m-p_j|}{\prod_{i=1}^n(p_i+m+1)\left(\sqrt{2m+1}\right)}$.

24. Suppose $\sum_{k=0}^n a_k t^k = 0$ for each $t \in (-\delta,\delta)$ where $a_k \in X$, a linear space. Show that each $a_k = 0$.

Chapter 5

Functions on Normed Linear Spaces

This chapter is about the general notion of functions defined on normed linear spaces even if the linear space is not finite dimensional.

5.1 $\mathscr{L}(V,W)$ as a Vector Space

In what follows, V, W will be vector spaces.

Definition 5.1.1 *The term $\mathscr{L}(V,W)$ signifies the set of linear maps from V to W. This means that for $v,u \in V$ and α,β scalars from $\mathbb{F}$, $L(\alpha u+\beta v) = \alpha L(u)+\beta L(v)$. Given $L,M \in \mathscr{L}(V,W)$ define a new element of $\mathscr{L}(V,W)$, denoted by $L+M$ according to the rule*[1] *$(L+M)v \equiv Lv+Mv$. For α a scalar and $L \in \mathscr{L}(V,W)$, define $\alpha L \in \mathscr{L}(V,W)$ by $\alpha L(v) \equiv \alpha(Lv)$.*

Note that if you have $V = \mathbb{R}^n$ and $W = \mathbb{R}^m$, an example of something in $\mathscr{L}(V,W)$ is given by $T\boldsymbol{v} \equiv A\boldsymbol{v}$ where A is a real $m \times n$ matrix.

You should verify that all the axioms of a vector space hold for $\mathscr{L}(V,W)$ with the above definitions of vector addition and scalar multiplication. What about the dimension of $\mathscr{L}(V,W)$?

Before answering this question, here is a useful lemma. It gives a way to define linear transformations and a way to tell when two of them are equal.

Lemma 5.1.2 *Let V and W be vector spaces and suppose $\{v_1,\cdots,v_n\}$ is a basis for V. Then if $L: V \to W$ is given by $Lv_k = w_k \in W$ and*

$$L\left(\sum_{k=1}^n a_k v_k\right) \equiv \sum_{k=1}^n a_k L v_k = \sum_{k=1}^n a_k w_k$$

then L is well defined and is in $\mathscr{L}(V,W)$. Also, if L,M are two linear transformations such that $Lv_k = Mv_k$ for all k, then $M = L$.

Proof: L is well defined on V because, since $\{v_1,\cdots,v_n\}$ is a basis, there is exactly one way to write a given vector of V as a linear combination. Next, observe that L is obviously linear from the definition. If L,M are equal on the basis, then if $\sum_{k=1}^n a_k v_k$ is an arbitrary vector of V,

$$L\left(\sum_{k=1}^n a_k v_k\right) = \sum_{k=1}^n a_k L v_k = \sum_{k=1}^n a_k M v_k = M\left(\sum_{k=1}^n a_k v_k\right)$$

and so $L = M$ because they give the same result for every vector in V. ■

The message is that when you define a linear transformation, it suffices to tell what it does to a basis.

Theorem 5.1.3 *Let V and W be finite dimensional linear spaces of dimension n and m respectively Then* $\dim(\mathscr{L}(V,W)) = mn$.

[1] Note that this is the standard way of defining the sum of two functions.

Proof: Let two sets of bases be $\{v_1,\cdots,v_n\}$ and $\{w_1,\cdots,w_m\}$ for V and W respectively. Using Lemma 5.1.2, let $w_iv_j \in \mathscr{L}(V,W)$ be the linear transformation defined on the basis, $\{v_1,\cdots,v_n\}$, by $w_iv_k(v_j) \equiv w_i\delta_{jk}$ where $\delta_{ik}=1$ if $i=k$ and 0 if $i\neq k$. I will show that $L\in\mathscr{L}(V,W)$ is a linear combination of these special linear transformations called dyadics.

Then let $L\in\mathscr{L}(V,W)$. Since $\{w_1,\cdots,w_m\}$ is a basis, there exist constants, d_{jk} such that $Lv_r=\sum_{j=1}^m d_{jr}w_j$ Now consider the following sum of dyadics. $\sum_{j=1}^m\sum_{i=1}^n d_{ji}w_jv_i$. Apply this to v_r. This yields

$$\sum_{j=1}^m\sum_{i=1}^n d_{ji}w_jv_i(v_r)=\sum_{j=1}^m\sum_{i=1}^n d_{ji}w_j\delta_{ir}=\sum_{j=1}^m d_{jr}w_i=Lv_r$$

Therefore, $L=\sum_{j=1}^m\sum_{i=1}^n d_{ji}w_jv_i$ showing the span of the dyadics is all of $\mathscr{L}(V,W)$.

Now consider whether these dyadics form a linearly independent set. Suppose

$$\sum_{i,k} d_{ik}w_iv_k=\mathbf{0}.$$

Are all the scalars d_{ik} equal to 0? $\mathbf{0}=\sum_{i,k}d_{ik}w_iv_k(v_l)=\sum_{i=1}^m d_{il}w_i$ so, since $\{w_1,\cdots,w_m\}$ is a basis, $d_{il}=0$ for each $i=1,\cdots,m$. Since l is arbitrary, this shows $d_{il}=0$ for all i and l. Thus these linear transformations form a basis and this shows that the dimension of $\mathscr{L}(V,W)$ is mn as claimed because there are m choices for the w_i and n choices for the v_j. ■

5.2 The Norm of a Linear Map, Operator Norm

Not surprisingly all of the above holds for a finite dimensional normed linear space. First here is an easy lemma which follows right away from Theorem 3.6.2, the theorem about equivalent formulations of continuity.

Lemma 5.2.1 *Let $(V,\|\cdot\|_V)$ and $(W,\|\cdot\|_W)$ be two normed linear spaces. Then a linear map $f:V\to W$ is continuous if and only if it takes bounded sets to bounded sets. (f is bounded) If V is finite dimensional, then f must be continuous.*

Proof: $\Longrightarrow$ Consider $f(B(0,1))$. If this is not bounded, then there exists $\|v^m\|_V\leq 1$ but $\|f(v^m)\|_W\geq m$. Then it follows that $\left\|f\left(\frac{v^m}{m}\right)\right\|_W\geq 1$ which is impossible for all m since $\left\|\frac{v^m}{m}\right\|\leq\frac{1}{m}$ and so continuity requires that $\lim_{m\to\infty}f\left(\frac{v^m}{m}\right)=0$ (Theorem 3.6.2). Thus there exists M such that $\|f(v)\|\leq M$ whenever $v\in B(0,1)$. In general, let S be a bounded set. Then $S\subseteq B(0,r)$ for large enough r. Hence, for $v\in B$, it follows that $v/2r\in B(0,1)$. It follows that $\|f(v/2r)\|_W\leq M$ and so $\|f(v)\|_W\leq 2rM$. Thus f takes bounded sets to bounded sets.

$\Longleftarrow$ Suppose f is bounded and not continuous. Then by Theorem 3.6.2 again, there is a sequence $v_n\to v$ but $f(v_n)$ fails to converge to $f(v)$. Then there exists $\varepsilon>0$ and a subsequence, still denoted as v_n such that $\|f(v_n)-f(v)\|=\|f(v_n-v)\|\geq\varepsilon$. Then

$$\left\|f\left(\frac{v_n-v}{\|v_n-v\|}\right)\right\|\geq\varepsilon\frac{1}{\|v_n-v\|}$$

The right side is unbounded, but the left is bounded, a contradiction.

Consider the last claim about continuity. Let $\{v_1,\cdots,v_n\}$ be a basis for V. By Lemma 4.4.7, if $y^m\to 0$, in V for $y^m=\sum_{k=1}^n y_k^m v_k$,then it follows that $\lim_{m\to\infty} y_k^m=0$ and consequently, $f(y^m)\to f(0)=0$. In general, if $y^m\to y$, then $(y^m-y)\to 0$ and so $f(y^m-y)=f(y^m)-f(y)\to 0$. That is, $f(y^m)\to f(y)$. ■

Definition 5.2.2 *For $f:(V,\|\cdot\|_V)\to(W,\|\cdot\|_W)$ continuous, it was just shown that there exists M such that $\|f(v)\|\le M,\ \ v\in B(0,1)$. It follows that, since $\frac{v}{2\|v\|}\in B(0,1)$, then $\|f(v)\|\le 2M\|v\|$. Therefore, letting $\|f\|\equiv \sup_{\|v\|\le 1}\|f(v)\|$ it follows that for all $v\in V$, $\|f(v)\|\le\|f\|\,\|v\|$ .Thus a linear map is bounded if and only if $\|f\|<\infty$ if and only if f is continuous. 'The number $\|f\|$ is called the operator norm. For X a real normed linear space, X' denotes the space $\mathscr{L}(X,\mathbb{R})$.*

You can show that for $\mathscr{L}(V,W)$ the space of bounded linear maps from V to W, $\mathscr{L}(V,W)$ becomes a normed linear space with this definition. This is true whether V,W are finite or infinite dimensional. You can also show that if W is complete then so is $\mathscr{L}(V,W)$. This is left as an exercise. Also, when the vector spaces are finite dimensional, Lemma 5.2.1 shows that any linear function f is automatically bounded, hence continuous, hence $\|f\|$ exists. Here is an interesting observation about the operator norm.

Lemma 5.2.3 *Let $f\in\mathscr{L}(V,W)$ and let $h\in\mathscr{L}(W,Z)$ where X,Y,Z are normed vector spaces. Then $\|h\circ f\|\le\|h\|\,\|f\|$.*

Proof: This follows right away from the definition. If $\|v\|\le 1$, then $\|f(v)\|\le\|f\|$. This explains the first inequality in the following.

$$\sup_{\|v\|\le 1}\|h\circ f(v)\|\le\sup_{\|w\|\le\|f\|}\|h(w)\|=\sup_{\|w\|\le\|f\|}\left\|h\left(\frac{w}{\|f\|}\right)\right\|\|f\|\le\|h\|\,\|f\|.\ ■$$

Theorem 5.2.4 *Let $(V,\|\cdot\|)$ be a normed linear space with basis $\{v_1,\cdots,v_n\}$ and field of scalars $\mathbb{F}$. Let $f:(\mathbb{F}^n,\|\cdot\|)\to(V,\|\cdot\|_V)$ be any linear map which is one to one and onto. Then both f and f^{-1} are continuous. Also the compact sets of $(V,\|\cdot\|_V)$ are exactly those which are closed and bounded.*

Proof: Define another norm $\|\cdot\|_1$ on $\mathbb{F}^n$ as follows. $\|\boldsymbol{x}\|_1\equiv\|f(\boldsymbol{x})\|_V$. Since f is one to one and onto and linear, this is indeed a norm. The details are left as an exercise. Then from the theorem on the equivalence of norms, there are positive constants δ,Δ such that $\delta\|\boldsymbol{x}\|\le\|f(\boldsymbol{x})\|_V\le\Delta\|\boldsymbol{x}\|$. Since f is one to one and onto, this implies $\delta\left\|f^{-1}(\boldsymbol{v})\right\|\le\|\boldsymbol{v}\|_V\le\Delta\left\|f^{-1}(\boldsymbol{v})\right\|$. The first of these above inequalities implies f is continuous. The second says $\left\|f^{-1}(\boldsymbol{v})\right\|\le\frac{1}{\delta}\|\boldsymbol{v}\|_V$ and so f^{-1} is continuous. Thus, from the above theorems, both f and f^{-1} map closed sets to closed sets, compact sets to compact sets, open sets to open sets and bounded sets to bounded sets.

Now let $K\subseteq V$ be closed and bounded. Then from the above observations, $f^{-1}(K)$ is also closed and bounded. Therefore, it is compact. Now $f\left(f^{-1}(K)\right)=K$ must be compact because the continuous image of a compact set is compact, Theorem 3.7.1. Conversely, if $K\subseteq V$ is compact, then by the theorem just mentioned, $f^{-1}(K)$ is compact and so it is closed and bounded. Hence $f\left(f^{-1}(K)\right)=K$ is also closed and bounded. ■

This is a remarkable theorem. It says that an algebraic isomorphism is also a homeomorphism which is what it means to say that the map takes open sets to open sets and the inverse does the same. In other words, there really isn't any algebraic or topological

distinction between a finite dimensional normed vector space of dimension n and $\mathbb{F}^n$. Of course when one considers geometry, this is not so.

Here is another interesting theorem about coordinate maps. It follows right away from earlier theorems.

Theorem 5.2.5 *Let $f:(V,\|\cdot\|_V)\to(W,\|\cdot\|_W)$ be a continuous function where here $(V,\|\cdot\|_V)$ is a normed linear space and $(W,\|\cdot\|_W)$ is a finite dimensional normed linear space with basis $\{w_1,\cdots,w_n\}$. Thus $f(v)\equiv\sum_{k=1}^n f_k(v)w_k$. Then f is continuous if and only if each f_k is a continuous $\mathbb{F}$ valued map.*

Proof: $\Longrightarrow$ First, why is f_k linear? This follows from

$$\sum_{k=1}^n(\alpha f_k(u)+\beta f_k(v))w_k=\alpha\sum_{k=1}^n f_k(u)w_k+\beta\sum_{k=1}^n f_k(v)w_k$$

$$=\alpha f(u)+\beta f(v)=f(\alpha u+\beta v)\equiv\sum_{k=1}^n f_k(\alpha u+\beta v)w_k$$

Why is the coordinate function f_k continuous? From Lemma 5.2.1, it suffices to verify that f_k is bounded. If this is not so, there exists $v_m,\|v_m\|_V\le 1$ but $|f_k(v_m)|_W\ge m$. It follows that $\left|f_k\left(\frac{v_m}{m}\right)\right|\ge 1$. Since f is continuous, and $v_m/m\to 0$, it follows that $f\left(\frac{v_m}{m}\right)\to 0$ in V. However, by Lemma 4.4.7, $f_k\left(\frac{v_m}{m}\right)\to 0$, a contradiction.

$\Longleftarrow$ If each coordinate function is continuous, then

$$\|f(v)-f(\hat{v})\|_W=\left\|\sum_{k=1}^n f_k(v)w_k-\sum_{k=1}^n f_k(\hat{v})w_k\right\|\le\sum_{k=1}^n|f_k(v)-f_k(\hat{v})|\,\|w_k\|_W$$

Since each f_k is continuous, this shows that f is also. ■

5.3 Continuous Functions in Normed Linear Space

Of course not all functions are linear. Continuous functions have already been discussed in general metric space, but now there are other considerations to consider due to the algebra available in a normed linear space. The following theorem includes these kinds of considerations for functions having values in a normed linear space.

Theorem 5.3.1 *Let f,g be continuous functions defined on D, a metric space. Also let α,β be scalars. Then the following hold.*

1. *$\alpha f+\beta g$ is continuous.*
2. *If $(W,\|\cdot\|_W)$ is an inner product space, then (f,g) defined as $(f,g)(v)\equiv(f(v),g(v))$, then (f,g) is continuous.*
3. *If f has values in $\mathbb{F}$ and g has values in $(W,\|\cdot\|_W)$, then fg is continuous.*

Proof: Say $v_n\to v$. Then

$$\|(\alpha f+\beta g)(v_n)-(\alpha f+\beta g)(v)\|\le|\alpha|\,\|f(v_n)-f(v)\|+|\beta|\,\|g(v_n)-g(v)\|$$

and the right side converges to 0 as $n\to\infty$ so this shows 1.

This follows from an easy computation. From the Cauchy Schwarz inequality,

$$\begin{aligned} & |(f,g)(v)-(f,g)(\hat{v})| \\ \leq\ & |(f(v),g(v))-(f(v),g(\hat{v}))|+|(f(v),g(\hat{v}))-(f(\hat{v}),g(\hat{v}))| \\ & \leq \|g(v)-g(\hat{v})\|\,\|f(v)\|+\|f(v)-f(\hat{v})\|\,\|g(\hat{v})\| \end{aligned}$$

Now since g is continuous at v and so $\|g(v)-g(\hat{v})\|<1$ provided $d(v,\hat{v})$ is small enough. Thus

$$\|g(\hat{v})\| \leq \|g(v)\|+1.$$

Hence if $d(v,\hat{v})$ is small enough,

$$|(f,g)(v)-(f,g)(\hat{v})| \leq (\|g(v)\|+1)\,\|f(v)-f(\hat{v})\|+\|f(v)\|\,\|g(v)-g(\hat{v})\|$$

Thus, by continuity of f,g at v, if $d(v,\hat{v})$ is sufficiently small, the right side is less than ε and so $f\cdot g$ is continuous at v. This shows 2. The proof of 3. is just like this. ■

Of course there are other things like cross product and determinant and so forth which are defined in terms of the component functions of f. Then these things will be continuous by an application of Theorem 5.2.5.

5.4 Polynomials

For functions of one variable, the special kind of functions known as a polynomial has a corresponding version when one considers a function of many variables. This is found in the next definition.

Definition 5.4.1 *Let α be an n dimensional multi-index. The meaning of this term is that $\boldsymbol{\alpha}=(\alpha_1,\cdots,\alpha_n)$ where each α_i is a positive integer or zero. Also, let $|\boldsymbol{\alpha}|\equiv\sum_{i=1}^n|\alpha_i|$. Then $\boldsymbol{x}^\alpha$ means $\boldsymbol{x}^\alpha\equiv x_1^{\alpha_1}x_2^{\alpha_2}\cdots x_3^{\alpha_n}$ where each $x_j\in\mathbb{F}$. An n dimensional polynomial of degree m is a function of the form $p(\boldsymbol{x})=\sum_{|\alpha|\leq m}d_\alpha\boldsymbol{x}^\alpha$. where the d_α are complex or real numbers, more generally in some normed linear space X. Rational functions are defined as the quotient of two real or complex valued polynomials. Thus these functions are defined on $\mathbb{F}^n$.*

For example, $f(\boldsymbol{x})=x_1x_2^2+7x_3^4x_1$ is a polynomial of degree 5 and $\frac{x_1x_2^2+7x_3^4x_1+x_2^3}{4x_1^3x_2^2+7x_3^2x_1-x_2^3}$ is a rational function.

Note that in the case of a rational function, the domain of the function might not be all of $\mathbb{F}^n$. For example, if $f(\boldsymbol{x})=\frac{x_1x_2^2+7x_3^4x_1+x_2^3}{x_2^2+3x_1^2-4}$,the domain of f would be all complex numbers such that $x_2^2+3x_1^2\neq 4$.

By Theorem 3.6.2 all polynomials are continuous. To see this, note that the function, $\pi_k(\boldsymbol{x})\equiv x_k$ is a continuous function because of the inequality

$$|\pi_k(\boldsymbol{x})-\pi_k(\boldsymbol{y})|=|x_k-y_k|\leq|\boldsymbol{x}-\boldsymbol{y}|.$$

Polynomials are simple sums of scalars times products of these functions. Similarly, by this theorem, rational functions, quotients of polynomials, are continuous at points where the denominator is non zero. More generally, if V is a normed vector space, consider a V valued function of the form $\boldsymbol{f}(\boldsymbol{x})\equiv\sum_{|\alpha|\leq m}\boldsymbol{d}_\alpha\boldsymbol{x}^\alpha$ where $\boldsymbol{d}_\alpha\in V$, sort of a V valued

polynomial. Then such a function is continuous by application of Theorem 3.6.2 and the above observation about the continuity of the functions π_k.

Thus there are lots of examples of continuous functions. However, it is even better than the above discussion indicates. As in the case of a function of one variable, an arbitrary continuous function can typically be approximated uniformly by a polynomial. This is the n dimensional version of the Weierstrass approximation theorem.

5.5 Weierstrass Approximation Theorem

An arbitrary continuous function defined on an interval can be approximated uniformly by a polynomial, there exists a similar theorem which is just a generalization of this which will hold for continuous functions defined on a box or more generally a closed and bounded set. However, we will settle for the case of a box first. The proof is based on the following lemma.

Lemma 5.5.1 *The following estimate holds for $x \in [0,1]$ and $m \geq 2$.*

$$\sum_{k=0}^{m} \binom{m}{k} (k-mx)^2 x^k (1-x)^{m-k} \leq \frac{1}{4} m$$

Proof: First of all, from the binomial theorem,

$$\sum_{k=0}^{m} \binom{m}{k} \left(e^{t(k-mx)}\right) x^k (1-x)^{m-k} = e^{-tmx} \sum_{k=0}^{m} \binom{m}{k} \left(e^{tk}\right) x^k (1-x)^{m-k}$$

$$= e^{-tmx} \left(1-x+xe^t\right)^m = e^{-tmx} g(t)^m,\ g(0) = 1, g'(0) = g''(0) = x$$

Take a partial derivative with respect to t twice.

$$\sum_{k=0}^{m} \binom{m}{k} (k-mx)^2 e^{t(k-mx)} x^k (1-x)^{m-k}$$

$$\begin{aligned} = \quad & (mx)^2 e^{-tmx} g(t)^m + 2(-mx) e^{-tmx} m g(t)^{m-1} g'(t) \\ & + e^{-tmx} \left[m(m-1) g(t)^{m-2} g'(t)^2 + m g(t)^{m-1} g''(t) \right] \end{aligned}$$

Now let $t = 0$ and note that the right side is $m(x-x^2) \leq m/4$ for $x \in [0,1]$. Thus

$$\sum_{k=0}^{m} \binom{m}{k} (k-mx)^2 x^k (1-x)^{m-k} = mx - mx^2 \leq m/4 \ \blacksquare$$

With this preparation, here is the first version of the Weierstrass approximation theorem. I will allow f to have values in a complete, real or complex normed linear space. Thus, $f \in C([0,1];X)$ where X is a Banach space, Definition 4.3.7. Thus this is a function which is continuous with values in X as discussed earlier with metric spaces.

Theorem 5.5.2 *Let $f \in C([0,1];X)$ and let the norm on X be denoted by $\|\cdot\|$.*

$$p_m(x) \equiv \sum_{k=0}^{m} \binom{m}{k} x^k (1-x)^{m-k} f\left(\frac{k}{m}\right).$$

Then these polynomials having coefficients in X converge uniformly to f on $[0,1]$.

Proof: Let $\|f\|_\infty$ denote the largest value of $\|f(x)\|$. By uniform continuity of f, there exists a $\delta > 0$ such that if $|x-x'| < \delta$, then $\|f(x)-f(x')\| < \varepsilon/2$. By the binomial theorem,

$$\begin{aligned}\|p_m(x)-f(x)\| &\le \sum_{k=0}^{m}\binom{m}{k}x^k(1-x)^{m-k}\left\|f\left(\frac{k}{m}\right)-f(x)\right\| \\ &\le \sum_{\left|\frac{k}{m}-x\right|<\delta}\binom{m}{k}x^k(1-x)^{m-k}\left\|f\left(\frac{k}{m}\right)-f(x)\right\| + \\ &\quad 2\|f\|_\infty \sum_{\left|\frac{k}{m}-x\right|\ge\delta}\binom{m}{k}x^k(1-x)^{m-k}\end{aligned}$$

Therefore,

$$\le \sum_{k=0}^{m}\binom{m}{k}x^k(1-x)^{m-k}\frac{\varepsilon}{2} + 2\|f\|_\infty \sum_{(k-mx)^2\ge m^2\delta^2}\binom{m}{k}x^k(1-x)^{m-k}$$

$$\begin{aligned}&\le \frac{\varepsilon}{2} + 2\|f\|_\infty \frac{1}{m^2\delta^2}\sum_{k=0}^{m}\binom{m}{k}(k-mx)^2x^k(1-x)^{m-k} \\ &\le \frac{\varepsilon}{2} + 2\|f\|_\infty \frac{1}{4}m\frac{1}{\delta^2m^2} < \varepsilon\end{aligned}$$

provided m is large enough. Thus $\|p_m - f\|_\infty < \varepsilon$ when m is large enough. ■

Note that we do not need to have X be complete in order for this to hold. It would have sufficed to have simply let X be a normed linear space.

Corollary 5.5.3 *If $f \in C([a,b];X)$ where X is a normed linear space, then there exists a sequence of polynomials which converge uniformly to f on $[a,b]$.*

Proof: Let $l : [0,1] \to [a,b]$ be one to one, linear and onto. Then $f \circ l$ is continuous on $[0,1]$ and so if $\varepsilon > 0$ is given, there exists a polynomial p such that for all $x \in [0,1]$, $\|p(x) - f \circ l(x)\| < \varepsilon$. Therefore, letting $y = l(x)$, it follows that for all $y \in [a,b]$, $\left\|p\left(l^{-1}(y)\right) - f(y)\right\| < \varepsilon$. ■

The exact form of the polynomial is as follows.

$$p(x) = \sum_{k=0}^{m}\binom{m}{k}x^k(1-x)^{m-k}f\left(l\left(\frac{k}{m}\right)\right)$$

$$p\left(l^{-1}(y)\right) = \sum_{k=0}^{m}\binom{m}{k}\left(l^{-1}(y)\right)^k\left(1-l^{-1}(y)\right)^{m-k}f\left(l\left(\frac{k}{m}\right)\right) \tag{5.1}$$

As another corollary, here is the version which will be used in Stone's generalization later.

Corollary 5.5.4 *Let f be a continuous function defined on $[-M,M]$ with $f(0) = 0$. Then there is a sequence of polynomials $\{p_m\}$, $p_m(0) = 0$ and*

$$\lim_{m\to\infty}\|p_m - f\|_\infty = 0$$

Proof: From Corollary 5.5.3 there exists a sequence of polynomials $\{\widehat{p_m}\}$ such that $\|\widehat{p_m} - f\|_\infty \to 0$. Simply consider $p_m = \widehat{p_m} - \widehat{p_m}(0)$. ■

5.6 Functions of Many Variables

First note that if $h : K \times H \to \mathbb{R}$ is a real valued continuous function where K, H are compact sets in metric spaces,

$$\max_{x\in K} h(x,y) \geq h(x,y), \text{ so } \max_{y\in H}\max_{x\in K} h(x,y) \geq h(x,y)$$

which implies $\max_{y\in H}\max_{x\in K} h(x,y) \geq \max_{(x,y)\in K\times H} h(x,y)$. The other inequality is also obtained.

Let $\boldsymbol{f} \in C(R_p;X)$ where $R_p = [0,1]^p$. Then let $\hat{\boldsymbol{x}}_p \equiv (x_1,...,x_{p-1})$. By Theorem 5.5.2, if n is large enough,

$$\max_{x_p\in[0,1]} \left\| \sum_{k=0}^{n} \boldsymbol{f}\left(\cdot, \frac{k}{n}\right) \binom{n}{k} x_p^k (1-x_p)^{n-k} - \boldsymbol{f}(\cdot, x_p) \right\|_{C\left([0,1]^{p-1};X\right)} < \frac{\varepsilon}{2}$$

Now $\boldsymbol{f}\left(\cdot, \frac{k}{n}\right) \in C(R_{p-1};X)$ and so by induction, there is a polynomial $\boldsymbol{p}_k(\hat{\boldsymbol{x}}_p)$ such that

$$\max_{\hat{\boldsymbol{x}}_p\in R_{p-1}} \left\| \boldsymbol{p}_k(\hat{\boldsymbol{x}}_p) - \binom{n}{k} \boldsymbol{f}\left(\hat{\boldsymbol{x}}_p, \frac{k}{n}\right) \right\|_X < \frac{\varepsilon}{(n+1)2}$$

Thus, letting $\boldsymbol{p}(\boldsymbol{x}) \equiv \sum_{k=0}^{n} \boldsymbol{p}_k(\hat{\boldsymbol{x}}_p) x_p^k (1-x_p)^{n-k}$,

$$\|\boldsymbol{p} - \boldsymbol{f}\|_{C\left(R_p;X\right)} \leq \max_{x_p\in[0,1]} \max_{\hat{\boldsymbol{x}}_p\in R_{p-1}} \left\|\boldsymbol{p}(\hat{\boldsymbol{x}}_p, x_p) - \boldsymbol{f}(\hat{\boldsymbol{x}}_p, x_p)\right\|_X < \varepsilon$$

where $\boldsymbol{p}$ is a polynomial with coefficients in X.

In general, if $R_p \equiv \prod_{k=1}^{p} [a_k, b_k]$, note that there is a linear function $l_k : [0,1] \to [a_k, b_k]$ which is one to one and onto. Thus $\boldsymbol{l}(\boldsymbol{x}) \equiv (l_1(x_1), ..., l_p(x_p))$ is a one to one and onto map from $[0,1]^p$ to R_p and the above result can be applied to $\boldsymbol{f} \circ \boldsymbol{l}$ to obtain a polynomial $\boldsymbol{p}$ with $\|\boldsymbol{p} - \boldsymbol{f} \circ \boldsymbol{l}\|_{C([0,1]^p;X)} < \varepsilon$. Thus $\left\|\boldsymbol{p} \circ \boldsymbol{l}^{-1} - \boldsymbol{f}\right\|_{C\left(R_p;X\right)} < \varepsilon$ and $\boldsymbol{p} \circ \boldsymbol{l}^{-1}$ is a polynomial. This proves the following theorem.

Theorem 5.6.1 *Let $\boldsymbol{f}$ be a function in $C(R;X)$ for X a normed linear space where $R \equiv \prod_{k=1}^{p} [a_k, b_k]$. Then for any $\varepsilon > 0$ there exists a polynomial $\boldsymbol{p}$ having coefficients in X such that $\|\boldsymbol{p} - \boldsymbol{f}\|_{C(R;X)} < \varepsilon$.*

These Bernstein polynomials are very remarkable approximations. It turns out that if f is $C^1([0,1];X)$, then $\lim_{n\to\infty} p_n'(x) \to f'(x)$ uniformly on $[0,1]$. This all works for functions of many variables as well, but here I will only show it for functions of one variable. I assume the reader knows about the derivative of a function of one variable. If not, skip this till later till after the derivative has been presented.

Lemma 5.6.2 *Let $f \in C^1([0,1])$ and let*

$$p_m(x) \equiv \sum_{k=0}^{m} \binom{m}{k} x^k (1-x)^{m-k} f\left(\frac{k}{m}\right)$$

be the m^{th} Bernstein polynomial. Then in addition to $\|p_m - f\|_{[0,1]} \to 0$, it also follows that

$$\left\|p_m' - f'\right\|_{[0,1]} \to 0$$

Proof: From simple computations,

$$\begin{aligned} p_m'(x) &= \sum_{k=1}^{m} \binom{m}{k} k x^{k-1} (1-x)^{m-k} f\left(\frac{k}{m}\right) \\ &\quad - \sum_{k=0}^{m-1} \binom{m}{k} x^k (m-k) (1-x)^{m-1-k} f\left(\frac{k}{m}\right) \\ &= \sum_{k=1}^{m} \frac{m(m-1)!}{(m-k)!(k-1)!} x^{k-1} (1-x)^{m-k} f\left(\frac{k}{m}\right) \\ &\quad - \sum_{k=0}^{m-1} \binom{m}{k} x^k (m-k) (1-x)^{m-1-k} f\left(\frac{k}{m}\right) \\ &= \sum_{k=0}^{m-1} \frac{m(m-1)!}{(m-1-k)!k!} x^k (1-x)^{m-1-k} f\left(\frac{k+1}{m}\right) \\ &\quad - \sum_{k=0}^{m-1} \frac{m(m-1)!}{(m-1-k)!k!} x^k (1-x)^{m-1-k} f\left(\frac{k}{m}\right) \\ &= \sum_{k=0}^{m-1} \frac{m(m-1)!}{(m-1-k)!k!} x^k (1-x)^{m-1-k} \left(f\left(\frac{k+1}{m}\right) - f\left(\frac{k}{m}\right) \right) \\ &= \sum_{k=0}^{m-1} \binom{m-1}{k} x^k (1-x)^{m-1-k} \left(\frac{f\left(\frac{k+1}{m}\right) - f\left(\frac{k}{m}\right)}{1/m} \right) \end{aligned}$$

By the mean value theorem,

$$\frac{f\left(\frac{k+1}{m}\right) - f\left(\frac{k}{m}\right)}{1/m} = f'\left(x_{k,m}\right),\ x_{k,m} \in \left(\frac{k}{m}, \frac{k+1}{m}\right)$$

Now the desired result follows as before from the uniform continuity of f' on $[0,1]$. Let $\delta > 0$ be such that if

$$|x-y| < \delta, \text{ then } \left|f'(x) - f'(y)\right| < \varepsilon$$

and let m be so large that $1/m < \delta/2$. Then if $\left|x - \frac{k}{m}\right| < \delta/2$, it follows that $\left|x - x_{k,m}\right| < \delta$ and so

$$\left|f'(x) - f'\left(x_{k,m}\right)\right| = \left| f'(x) - \frac{f\left(\frac{k+1}{m}\right) - f\left(\frac{k}{m}\right)}{1/m} \right| < \varepsilon.$$

Now as before, letting $M \geq |f'(x)|$ for all x,

$$\begin{aligned} \left|p_m'(x) - f'(x)\right| &\leq \sum_{k=0}^{m-1} \binom{m-1}{k} x^k (1-x)^{m-1-k} \left|f'\left(x_{k,m}\right) - f'(x)\right| \\ &\leq \sum_{\left\{x : \left|x - \frac{k}{m}\right| < \frac{\delta}{2}\right\}} \binom{m-1}{k} x^k (1-x)^{m-1-k} \varepsilon \\ &\quad + M \sum_{k=0}^{m-1} \binom{m-1}{k} \frac{4(k-mx)^2}{m^2\delta^2} x^k (1-x)^{m-1-k} \end{aligned}$$

$$\leq \varepsilon + 4M\frac{1}{4}m\frac{1}{m^2\delta^2} = \varepsilon + M\frac{1}{m\delta^2} < 2\varepsilon$$

whenever m is large enough. Thus this proves uniform convergence. ■

There is a more general version of the Weierstrass theorem which is easy to get. It depends on the Tietze extension theorem, a wonderful little result which is interesting for its own sake.

5.7 A Generalization

This is an interesting theorem which holds in arbitrary normal topological spaces. In particular it holds in metric space and this is the context in which it will be discussed. First, review Lemma 3.12.1.

Lemma 5.7.1 *Let H, K be two nonempty disjoint closed subsets of X. Then there exists a continuous function, $g : X \to [-1/3, 1/3]$ such that $g(H) = -1/3$, $g(K) = 1/3, g(X) \subseteq [-1/3, 1/3]$.*

Proof: Let $f(\boldsymbol{x}) \equiv \frac{\text{dist}(\boldsymbol{x},H)}{\text{dist}(\boldsymbol{x},H)+\text{dist}(\boldsymbol{x},K)}$. The denominator is never equal to zero because if $\text{dist}(\boldsymbol{x},H) = 0$, then $\boldsymbol{x} \in H$ because H is closed. (To see this, pick $\boldsymbol{h}_k \in B(\boldsymbol{x}, 1/k) \cap H$. Then $\boldsymbol{h}_k \to \boldsymbol{x}$ and since H is closed, $\boldsymbol{x} \in H$.) Similarly, if $\text{dist}(\boldsymbol{x},K) = 0$, then $\boldsymbol{x} \in K$ and so the denominator is never zero as claimed. Hence f is continuous and from its definition, $f = 0$ on H and $f = 1$ on K. Now let $g(\boldsymbol{x}) \equiv \frac{2}{3}\left(f(\boldsymbol{x}) - \frac{1}{2}\right)$. Then g has the desired properties. ■

Definition 5.7.2 *For $f : M \subseteq X \to \mathbb{R}$, let $\|f\|_M \equiv \sup\{|f(\boldsymbol{x})| : \boldsymbol{x} \in M\}$. This is just notation. I am not claiming this is a norm.*

Lemma 5.7.3 *Suppose M is a closed set in X and suppose $f : M \to [-1, 1]$ is continuous at every point of M. Then there exists a function, g which is defined and continuous on all of X such that $\|f - g\|_M \leq \frac{2}{3}$, $g(X) \subseteq [-1/3, 1/3]$. If X is a normed vector space,and f is odd, meaning that M is symmetric ($\boldsymbol{x} \in M$ if and only if $-\boldsymbol{x} \in M$) and $f(-\boldsymbol{x}) = -f(\boldsymbol{x})$. Then we can assume g is also odd.*

Proof: Let $H = f^{-1}([-1, -1/3]), K = f^{-1}([1/3, 1])$. Thus H and K are disjoint closed subsets of M. Suppose first H, K are both nonempty. Then by Lemma 5.7.1 there exists g such that g is a continuous function defined on all of X and $g(H) = -1/3$, $g(K) = 1/3$, and $g(X) \subseteq [-1/3, 1/3]$. It follows $\|f - g\|_M < 2/3$. If $H = \emptyset$, then f has all its values in $[-1/3, 1]$ and so letting $g \equiv 1/3$, the desired condition is obtained. If $K = \emptyset$, let $g \equiv -1/3$. If both $H, K = \emptyset$, let $g = 0$.

When M is symmetric and f is odd, $g(\boldsymbol{x}) \equiv \frac{1}{3}\frac{\text{dist}(\boldsymbol{x},H)-\text{dist}(\boldsymbol{x},K)}{\text{dist}(\boldsymbol{x},H)+\text{dist}(\boldsymbol{x},K)}$. When $\boldsymbol{x} \in H$ this gives $\frac{1}{3}\frac{-\text{dist}(\boldsymbol{x},K)}{\text{dist}(\boldsymbol{x},K)} = -\frac{1}{3}$. Then $\boldsymbol{x} \in K$, this gives $\frac{1}{3}\frac{\text{dist}(\boldsymbol{x},H)}{\text{dist}(\boldsymbol{x},H)} = \frac{1}{3}$. Also $g(H) = -1/3$, $f(H) \subseteq [-1, -1/3]$ so for $\boldsymbol{x} \in H, |g(\boldsymbol{x}) - f(\boldsymbol{x})| \leq \frac{2}{3}$. It is similar for $\boldsymbol{x} \in K$. If $\boldsymbol{x}$ is in neither H nor K, then $g(\boldsymbol{x}) \in [-1/3, 1/3]$ and so is $f(\boldsymbol{x})$. Thus $\|f - g\|_M \leq \frac{2}{3}$. Now by assumption, since f is odd, $H = -K$. It is clear that g is odd because

$$\begin{aligned} g(-\boldsymbol{x}) &= \frac{1}{3}\frac{\text{dist}(-\boldsymbol{x},H) - \text{dist}(-\boldsymbol{x},K)}{\text{dist}(-\boldsymbol{x},H) + \text{dist}(-\boldsymbol{x},K)} = \frac{1}{3}\frac{\text{dist}(-\boldsymbol{x},-K) - \text{dist}(-\boldsymbol{x},-H)}{\text{dist}(-\boldsymbol{x},-K) + \text{dist}(-\boldsymbol{x},-H)} \\ &= \frac{1}{3}\frac{\text{dist}(\boldsymbol{x},K) - \text{dist}(\boldsymbol{x},H)}{\text{dist}(\boldsymbol{x},K) + \text{dist}(\boldsymbol{x},H)} = -g(\boldsymbol{x}). \ ■ \end{aligned}$$

Lemma 5.7.4 *Suppose M is a closed set in X and suppose $f : M \to [-1,1]$ is continuous at every point of M. Then there exists a function g which is defined and continuous on all of X such that $g = f$ on M and g has its values in $[-1,1]$. If X is a normed linear space and f is odd, then we can also assume g is odd.*

Proof: Using Lemma 5.7.3, let g_1 be such that $g_1(X) \subseteq [-1/3, 1/3]$ and $\|f - g_1\|_M \leq \frac{2}{3}$. Suppose $g_1, \cdots, g_m$ have been chosen such that $g_j(X) \subseteq [-1/3, 1/3]$ and

$$\left\| f - \sum_{i=1}^{m} \left(\frac{2}{3}\right)^{i-1} g_i \right\|_M < \left(\frac{2}{3}\right)^m. \tag{5.2}$$

This has been done for $m = 1$. Then $\left\| \left(\frac{3}{2}\right)^m \left(f - \sum_{i=1}^m \left(\frac{2}{3}\right)^{i-1} g_i \right) \right\|_M \leq 1$ and so

$$\left(\frac{3}{2}\right)^m \left(f - \sum_{i=1}^{m} \left(\frac{2}{3}\right)^{i-1} g_i \right)$$

can play the role of f in the first step of the proof. Therefore, there exists g_{m+1} defined and continuous on all of X such that its values are in $[-1/3, 1/3]$ and

$$\left\| \left(\frac{3}{2}\right)^m \left(f - \sum_{i=1}^{m} \left(\frac{2}{3}\right)^{i-1} g_i \right) - g_{m+1} \right\|_M \leq \frac{2}{3}.$$

Hence

$$\left\| \left(f - \sum_{i=1}^{m} \left(\frac{2}{3}\right)^{i-1} g_i \right) - \left(\frac{2}{3}\right)^m g_{m+1} \right\|_M \leq \left(\frac{2}{3}\right)^{m+1}.$$

It follows there exists a sequence, $\{g_i\}$ such that each has its values in $[-1/3, 1/3]$ and for every m 5.2 holds. Then let $g(\boldsymbol{x}) \equiv \sum_{i=1}^{\infty} \left(\frac{2}{3}\right)^{i-1} g_i(\boldsymbol{x})$. It follows

$$|g(\boldsymbol{x})| \leq \left| \sum_{i=1}^{\infty} \left(\frac{2}{3}\right)^{i-1} g_i(\boldsymbol{x}) \right| \leq \sum_{i=1}^{m} \left(\frac{2}{3}\right)^{i-1} \frac{1}{3} \leq 1$$

and $\left| \left(\frac{2}{3}\right)^{i-1} g_i(\boldsymbol{x}) \right| \leq \left(\frac{2}{3}\right)^{i-1} \frac{1}{3}$ so the Weierstrass M test applies and shows convergence is uniform. Therefore g must be continuous by Theorem 3.9.3. The estimate 5.2 implies $f = g$ on M. The last claim follows because we can take each g_i odd. ■

The following is the Tietze extension theorem.

Theorem 5.7.5 *Let M be a closed nonempty subset of a metric space X and let $f : M \to [a,b]$ be continuous at every point of M. Then there exists a function g continuous on all of X which coincides with f on M such that $g(X) \subseteq [a,b]$. If $[a,b]$ is centered on 0, and if X is a normed linear space and f is odd, then we can obtain that g is also odd.*

Proof: Let $f_1(\boldsymbol{x}) = 1 + \frac{2}{b-a}(f(\boldsymbol{x}) - b)$. Then f_1 satisfies the conditions of Lemma 5.7.4 and so there exists $g_1 : X \to [-1,1]$ such that g is continuous on X and equals f_1 on M. Let $g(\boldsymbol{x}) = (g_1(\boldsymbol{x}) - 1)\left(\frac{b-a}{2}\right) + b$. This works. The last claim follows from the same arguments which gave Lemma 5.7.4 or the change of variables just given. ■

Corollary 5.7.6 *Let M be a closed nonempty subset of a metric space X and let $f : M \to [a,b]$ be continuous at every point of M. Also let $\|f-g\| \leq \varepsilon$. Then there exists continuous $\hat{f}$ extending f with $\hat{f}(X) \subseteq [a,b]$ and $\hat{g}$ extending g such that $\hat{g}(X) \subseteq [a-\varepsilon, b+\varepsilon]$. Also $\left\|\hat{f}-\hat{g}\right\| \leq \varepsilon$.*

Proof: Let $\hat{f}$ be the extension of f from the above theorem. Now let F be the extension of $f-g$ with $\|F\| \leq \varepsilon$. Then let $\hat{g} = \hat{f} - F$. Then for $x \in M, \hat{g}(x) = f(x) - (f(x) - g(x)) = g(x)$. Thus it extends g and clearly $\hat{g}(X) \subseteq [a-\varepsilon, b+\varepsilon]$. ■

With the Tietze extension theorem, here is a better version of the Weierstrass approximation theorem.

Theorem 5.7.7 *Let K be a closed and bounded subset of $\mathbb{R}^p$ and let $f : K \to \mathbb{R}$ be continuous. Then there exists a sequence of polynomials $\{p_m\}$ such that*

$$\lim_{m\to\infty} (\sup\{|f(\boldsymbol{x}) - p_m(\boldsymbol{x})| : \boldsymbol{x} \in K\}) = 0.$$

In other words, the sequence of polynomials converges uniformly to f on K.

Proof: By the Tietze extension theorem, there exists an extension of f to a continuous function g defined on all $\mathbb{R}^p$ such that $g = f$ on K. Now since K is bounded, there exist intervals, $[a_k, b_k]$ such that $K \subseteq \prod_{k=1}^p [a_k, b_k] = R$. Then by the Weierstrass approximation theorem, Theorem 5.6.1 there exists a sequence of polynomials $\{p_m\}$ converging uniformly to g on R. Therefore, this sequence of polynomials converges uniformly to $g = f$ on K as well. This proves the theorem. ■

By considering the real and imaginary parts of a function which has values in $\mathbb{C}$ one can generalize the above theorem.

Corollary 5.7.8 *Let K be a closed and bounded subset of $\mathbb{R}^p$ and let $f : K \to \mathbb{F}$ be continuous. Then there exists a sequence of polynomials $\{p_m\}$ such that*

$$\lim_{m\to\infty} (\sup\{|f(\boldsymbol{x}) - p_m(\boldsymbol{x})| : \boldsymbol{x} \in K\}) = 0.$$

In other words, the sequence of polynomials converges uniformly to f on K.

More generally, the function f could have values in $\mathbb{R}^p$. There is no change in the proof. You just use norm symbols rather than absolute values and nothing at all changes in the theorem where the function is defined on a rectangle. Then you apply the Tietze extension theorem to each component in the case the function has values in $\mathbb{R}^p$.

5.8 An Approach to the Integral

With the Weierstrass approximation theorem, you can give a rigorous definition of the Riemann integral without wading in to Riemann sums. This shows the integral can be defined directly from very simple ideas. First is a short review of the derivative of a function of one variable.

Definition 5.8.1 *Let $f : [a,b] \to \mathbb{R}$. Then $f'(x) \equiv \lim_{x\to 0} \frac{f(x+h)-f(x)}{h}$ where h is always such that $x, x+h$ are both in the interval $[a,b]$ so we include derivatives at the right and left end points in this definition.*

The most important theorem about derivatives of functions of one variable is the mean value theorem.

Theorem 5.8.2 *Let $f:[a,b]\to\mathbb{R}$ be continuous. Then if the maximum value of f occurs at a point $x\in(a,b)$, it follows that if $f'(x)=0$. If f achieves a minimum at $x\in(a,b)$ where $f'(x)$ exists, it also follows that $f'(x)=0$.*

Proof: By Theorem 3.7.2, f achieves a maximum at some point x. If $f'(x)$ exists, then

$$f'(x)=\lim_{h\to 0+}\frac{f(x+h)-f(x)}{h}=\lim_{h\to 0-}\frac{f(x+h)-f(x)}{h}$$

However, the first limit is non-positive while the second is non-negative and so $f'(x)=0$. The situation is similar if the minimum occurs at $x\in(a,b)$. ■

The Cauchy mean value theorem follows. The usual one is obtained by letting $g(x)=x$.

Theorem 5.8.3 *Let f,g be continuous on $[a,b]$ and differentiable on (a,b). Then there exists $x\in(a,b)$ such that $f'(x)(g(b)-g(a))=g'(x)(f(b)-f(a))$. If $g(x)=x$, this yields $f(b)-f(a)=f'(x)(b-a)$, also $f(a)-f(b)=f'(x)(a-b)$.*

Proof: Let $h(x)\equiv f(x)(g(b)-g(a))-g(x)(f(b)-f(a))$. Then

$$h(a)=h(b)=f(a)g(b)-g(a)f(b).$$

If h is constant, then pick any $x\in(a,b)$ and $h'(x)=0$. If h is not constant, then it has either a maximum or a minimum on (a,b) and so if x is the point where either occurs, then $h'(x)=0$ which proves the theorem. ■

Recall that an antiderivative of a function f is just a function F such that $F'=f$. You know how to find an antiderivative for a polynomial. $\left(\frac{x^{n+1}}{n+1}\right)'=x^n$ so $\int\sum_{k=1}^n a_k x^k=\sum_{k=1}^n a_k\frac{x^{k+1}}{k+1}+C$. With this information and the Weierstrass theorem, it is easy to define integrals of continuous functions with all the properties presented in elementary calculus courses. It is an approach which does not depend on Riemann sums yet still gives the fundamental theorem of calculus. Note that if $F'(x)=0$ for x in an interval, then for x,y in that interval, $F(y)-F(x)=0(y-x)$ so F is a constant. Thus, if $F'=G'$ on an open interval, F,G continuous on the closed interval, it follows that $F-G$ is a constant and so $F(b)-F(a)=G(b)-G(a)$.

Definition 5.8.4 *For $p(x)$ a polynomial on $[a,b]$, let $P'(x)=p(x)$. Thus, by the mean value theorem if $P',\hat{P}'$ both equal p, it follows that $P(b)-P(a)=\hat{P}(b)-\hat{P}(a)$. Then define $\int_a^b p(x)\,dx\equiv P(b)-P(a)$. If $f\in C([a,b])$, define $\int_a^b f(x)\,dx\equiv\lim_{n\to\infty}\int_a^b p_n(x)\,dx$ where*

$$\lim_{n\to\infty}\|p_n-f\|\equiv\lim_{n\to\infty}\max_{x\in[a,b]}|f(x)-p_n(x)|=0$$

Proposition 5.8.5 *The above integral is well defined and satisfies the following properties.*

1. *$\int_a^b f\,dx=f(\hat{x})(b-a)$ for some $\hat{x}$ between a and b. Thus $\left|\int_a^b f\,dx\right|\le\|f\|\,|b-a|$.*

2. *If f is continuous on an interval which contains all necessary intervals,*

$$\int_a^c f dx + \int_c^b f dx = \int_a^b f dx, \text{ so } \int_a^b f dx + \int_b^a f dx = \int_b^b f dx = 0$$

3. *If $F(x) \equiv \int_a^x f dt$, Then $F'(x) = f(x)$ so any continuous function has an antiderivative, and for any $a \neq b$, $\int_a^b f dx = G(b) - G(a)$ whenever $G' = f$ on the open interval determined by a, b and G continuous on the closed interval determined by a, b. Also,*

$$\int_a^b (\alpha f(x) + \beta g(x)) dx = \alpha \int_a^b f(x) dx + \beta \int_a^b \beta g(x) dx$$

If $a < b$, and $f(x) \geq 0$, then $\int_a^b f dx \geq 0$. Also $\left| \int_a^b f dx \right| \leq \left| \int_a^b |f| dx \right|$.

4. $\int_a^b 1 dx = b - a$.

Proof: First, why is the integral well defined? With notation as in the above definition, the mean value theorem implies

$$\int_a^b p(x) dx \equiv P(b) - P(a) = p(\hat{x})(b-a) \tag{5.3}$$

where $\hat{x}$ is between a and b and so $\left| \int_a^b p(x) dx \right| \leq \|p\| |b-a|$. If $\|p_n - f\| \to 0$, then $\lim_{m,n\to\infty} \|p_n - p_m\| = 0$ and so

$$\left| \int_a^b p_n(x) dx - \int_a^b p_m(x) dx \right| = |(P_n(b) - P_n(a)) - (P_m(b) - P_m(a))|$$

$$= |(P_n(b) - P_m(b)) - (P_n(a) - P_m(a))| = \left| \int_a^b (p_n - p_m) dx \right| \leq \|p_n - p_m\| |b-a|$$

Thus the limit exists because $\left\{ \int_a^b p_n dx \right\}_n$ is a Cauchy sequence and $\mathbb{R}$ is complete.

From 5.3, 1. holds for a polynomial $p(x)$. Let $\|p_n - f\| \to 0$. Then by definition,

$$\int_a^b f dx \equiv \lim_{n\to\infty} \int_a^b p_n dx = p_n(x_n)(b-a) \tag{5.4}$$

for some x_n in the open interval determined by (a,b). By compactness, there is a further subsequence, still denoted with n such that $x_n \to x \in [a,b]$. Then fixing m such that $\|f - p_n\| < \varepsilon$ whenever $n \geq m$, assume $n > m$. Then $\|p_m - p_n\| \leq \|p_m - f\| + \|f - p_n\| < 2\varepsilon$ and so

$$|f(x) - p_n(x_n)| \leq |f(x) - f(x_n)| + |f(x_n) - p_m(x_n)| + |p_m(x_n) - p_n(x_n)|$$

$$\leq |f(x) - f(x_n)| + \|f - p_m\| + \|p_m - p_n\| < |f(x) - f(x_n)| + 3\varepsilon$$

Now if n is still larger, continuity of f shows that $|f(x) - p_n(x_n)| < 4\varepsilon$. Since ε is arbitrary, $p_n(x_n) \to f(x)$ and so, passing to the limit with this subsequence in 5.4 yields 1.

Now consider 2. It holds for polynomials $p(x)$ obviously. So let $\|p_n - f\| \to 0$. Then

$$\int_a^c p_n dx + \int_c^b p_n dx = \int_a^b p_n dx$$

Pass to a limit as $n \to \infty$ and use the definition to get 2. Also note that $\int_b^b f(x)\,dx = 0$ follows from the definition.

Next consider 3. Let $h \neq 0$ and let x be in the open interval determined by a and b. Then for small h,

$$\frac{F(x+h)-F(x)}{h} = \frac{1}{h}\int_x^{x+h} f(t)\,dt = f(x_h)$$

where x_h is between x and $x+h$. Let $h \to 0$. By continuity of f, it follows that the limit of the right side exists and so

$$\lim_{h\to 0}\frac{F(x+h)-F(x)}{h} = \lim_{h\to 0} f(x_h) = f(x)$$

If x is either end point, the argument is the same except you have to pay attention to the sign of h so that both x and $x+h$ are in $[a,b]$. Thus F is continuous on $[a,b]$ and F' exists on (a,b) so if G is an antiderivative,

$$\int_a^b f(t)\,dt \equiv F(b) = F(b)-F(a) = G(b)-G(a)$$

The claim that the integral is linear is obvious from this. Indeed, if $F' = f, G' = g$,

$$\begin{aligned}\int_a^b (\alpha f(t)+\beta g(t))\,dt &= \alpha F(b)+\beta G(b)-(\alpha F(a)+\beta G(a))\\ &= \alpha(F(b)-F(a))+\beta(G(b)-G(a))\\ &= \alpha\int_a^b f(t)\,dt+\beta\int_a^b g(t)\,dt\end{aligned}$$

If $f \geq 0$, then the mean value theorem implies that for some

$$t \in (a,b), F(b)-F(a) = \int_a^b f dx = f(t)(b-a) \geq 0.$$

Thus

$$\int_a^b (|f|-f)\,dx \geq 0, \int_a^b (|f|+f)\,dx \geq 0$$

and so $\int_a^b |f|\,dx \geq \int_a^b f dx$ and $\int_a^b |f|\,dx \geq -\int_a^b f dx$ so this proves $\left|\int_a^b f dx\right| \leq \int_a^b |f|\,dx$. This, along with part 2 implies the other claim that $\left|\int_a^b f dx\right| \leq \left|\int_a^b |f|\,dx\right|$.

The last claim is obvious because an antiderivative of 1 is $F(x) = x$. ■

Note also that the usual change of variables theorem is available because if $F' = f$, then $f(g(x))g'(x) = \frac{d}{dx}F(g(x))$ so that, from the above proposition,

$$F(g(b))-F(g(a)) = \int_{g(a)}^{g(b)} f(y)\,dy = \int_a^b f(g(x))g'(x)\,dx.$$

We usually let $y = g(x)$ and $dy = g'(x)\,dx$ and then change the limits as indicated above, equivalently we massage the expression to look like the above. Integration by parts also follows from differentiation rules.

Consider the iterated integral $\int_{a_1}^{b_1}\cdots\int_{a_p}^{b_p} \alpha x_1^{\alpha_1}\cdots x_p^{\alpha_p} dx_p\cdots dx_1$. It means just what it meant in calculus. You do the integral with respect to x_p first, keeping the other variables

constant, obtaining a polynomial function of the other variables. Then you do this one with respect to x_{p-1} and so forth. Thus, doing the computation, it reduces to

$$\alpha\prod_{k=1}^{p}\left(\int_{a_k}^{b_k}x_k^{\alpha_k}dx_k\right)=\alpha\prod_{k=1}^{p}\left(\frac{b^{\alpha_k+1}}{\alpha_k+1}-\frac{a^{\alpha_k+1}}{\alpha_k+1}\right)$$

and the same thing would be obtained for any other order of the iterated integrals. Since each of these integrals is linear, it follows that if $(i_1,\cdots,i_p)$ is any permutation of $(1,\cdots,p)$, then for any polynomial q,

$$\int_{a_1}^{b_1}\cdots\int_{a_p}^{b_p}q(x_1,...,x_p)\,dx_p\cdots dx_1=\int_{a_{i_p}}^{b_{i_1}}\cdots\int_{a_{i_p}}^{b_{i_p}}q(x_1,...,x_p)\,dx_{i_p}\cdots dx_{i_1}$$

Now let $f:\prod_{k=1}^{p}[a_k,b_k]\to\mathbb{R}$ be continuous. Then each iterated integral results in a continuous function of the remaining variables and so the iterated integral makes sense. For example, by Proposition 5.8.5, $\left|\int_c^d f(x,y)\,dy-\int_c^d f(\hat{x},y)\,dy\right|=$

$$\left|\int_c^d\left(f(x,y)-f(\hat{x},y)\right)dy\right|\leq\max_{y\in[c,d]}|f(x,y)-f(\hat{x},y)|<\varepsilon$$

if $|x-\hat{x}|$ is sufficiently small, thanks to uniform continuity of f on the compact set $[a,b]\times[c,d]$. Thus it makes perfect sense to consider the iterated integral $\int_a^b\int_c^d f(x,y)\,dydx$. Then using Proposition 5.8.5 on the iterated integrals along with Theorem 5.6.1, there exists a sequence of polynomials which converges to f uniformly $\{p_n\}$. Then applying Proposition 5.8.5 repeatedly,

$$\left|\int_{a_{i_p}}^{b_{i_1}}\cdots\int_{a_{i_p}}^{b_{i_p}}f(\boldsymbol{x})\,dx_p\cdots dx_1-\int_{a_{i_p}}^{b_{i_1}}\cdots\int_{a_{i_p}}^{b_{i_p}}p_n(\boldsymbol{x})\,dx_p\cdots dx_1\right|$$

$$\leq\|f-p_n\|\prod_{k=1}^{p}|b_k-a_k| \tag{5.5}$$

With this, it is easy to prove a rudimentary Fubini theorem valid for continuous functions.

Theorem 5.8.6 *$f:\prod_{k=1}^{p}[a_k,b_k]\to\mathbb{R}$ be continuous. Then for $(i_1,\cdots,i_p)$ any permutation of $(1,\cdots,p)$,*

$$\int_{a_{i_p}}^{b_{i_1}}\cdots\int_{a_{i_p}}^{b_{i_p}}f(\boldsymbol{x})\,dx_{i_p}\cdots dx_{i_1}=\int_{a_1}^{b_1}\cdots\int_{a_p}^{b_p}f(\boldsymbol{x})\,dx_p\cdots dx_1$$

If $f\geq 0$, then the iterated integrals are nonnegative if each $a_k\leq b_k$.

Proof: Let $\|p_n-f\|\to 0$ where p_n is a polynomial. Then from 5.5,

$$\int_{a_{i_1}}^{b_{i_1}}\cdots\int_{a_{i_p}}^{b_{i_p}}f(\boldsymbol{x})\,dx_{i_p}\cdots dx_{i_1}=\lim_{n\to\infty}\int_{a_{i_p}}^{b_{i_1}}\cdots\int_{a_{i_p}}^{b_{i_p}}p_n(\boldsymbol{x})\,dx_{i_p}\cdots dx_{i_1}$$

$$=\lim_{n\to\infty}\int_{a_1}^{b_1}\cdots\int_{a_p}^{b_p}p_n(\boldsymbol{x})\,dx_p\cdots dx_1=\int_{a_1}^{b_1}\cdots\int_{a_p}^{b_p}f(\boldsymbol{x})\,dx_p\cdots dx_1\ \blacksquare$$

You could replace f with $f\mathcal{X}_G$ where $\mathcal{X}_G(\boldsymbol{x})=1$ if $\boldsymbol{x}\in G$ and 0 otherwise provided each section of G consisting of holding all variables constant but 1, consists of finitely many intervals. Thus you can integrate over all the usual sets encountered in beginning calculus.

5.9 The Stone Weierstrass Approximation Theorem

There is a profound generalization of the Weierstrass approximation theorem due to Stone. It has to be one of the most elegant things available. It holds on locally compact Hausdorff spaces but here I will show the version which is valid on compact sets. Later the more general version is discussed.

Definition 5.9.1 *$\mathscr{A}$ is an algebra of functions if $\mathscr{A}$ is a vector space and if whenever $f,g \in \mathscr{A}$ then $fg \in \mathscr{A}$.*

To begin with assume that the field of scalars is $\mathbb{R}$. This will be generalized later. Theorem 5.5.2 implies the following corollary. See Corollary 5.5.3.

Corollary 5.9.2 *The polynomials are dense in $C([a,b])$.*

Here is another approach to proving this theorem. It is the original approach used by Weierstrass. Let $m \in \mathbb{N}$ and consider c_m such that $\int_{-1}^{1} c_m \left(1-x^2\right)^m dx = 1$. Then

$$1 = 2\int_0^1 c_m \left(1-x^2\right)^m dx \geq 2c_m \int_0^1 (1-x)^m dx = 2c_m \frac{1}{m+1}$$

so $c_m \leq m+1$. Then

$$\int_\delta^1 c_m \left(1-x^2\right)^m dx + \int_{-1}^{-\delta} c_m \left(1-x^2\right)^m dx \leq 2(m+1)\left(1-\delta^2\right)^m$$

which converges to 0. Thus

$$\lim_{m\to\infty} \sup_{x\notin[-\delta,\delta]} c_m \left(1-x^2\right)^m = 0 \tag{5.6}$$

Now let $\phi_n(t) \equiv c_m \left(1-t^2\right)^m$. Consider $f \in C([-1,1])$ and extend to let $f(x) = f(1)$ if $x > 1$ and $f(x) = f(-1)$ if $x < -1$ and define $p_m(x) \equiv \int_{-1}^{1} f(x-t)\,\phi_m(t)\,dt$. Then

$$|p_m(x) - f(x)| \leq \int_{-1}^{1} |f(x-t) - f(x)|\,\phi_m(t)\,dt$$

$$\begin{aligned} \leq & \int_{-1}^{1} \mathscr{X}_{[-\delta,\delta]}(t)\,|f(x-t) - f(x)|\,\phi_m(t)\,dt \\ & + \int_{-1}^{1} \mathscr{X}_{[-1,1]\setminus[-\delta,\delta]}(t)\,|f(x-t) - f(x)|\,\phi_m(t)\,dt \end{aligned}$$

Choose δ so small that if $|x-y| < \delta$, then $|f(x) - f(y)| < \varepsilon$. Also let $M \geq \max_x |f(x)|$. Then

$$\begin{aligned} |p_m(x) - f(x)| &\leq \varepsilon \int_{-1}^{1} \phi_m(t)\,dt + 2M \int_{-1}^{1} \mathscr{X}_{[-1,1]\setminus[-\delta,\delta]}(t)\,\phi_m(t)\,dt \\ &= \varepsilon + 2M \int_{-1}^{1} \mathscr{X}_{[-1,1]\setminus[-\delta,\delta]}(t)\,\phi_m(t)\,dt \end{aligned}$$

From 5.6, The second term is no larger than

$$2M \int_{-1}^{1} \mathscr{X}_{[-1,1]\setminus[-\delta,\delta]}(t)\,\varepsilon dt \leq 4M\varepsilon$$

whenever m is large enough. Hence, for large enough m,

$$\sup_{x\in[-1,1]} |p_m(x) - f(x)| \leq (1+4M)\,\varepsilon$$

Since ε is arbitrary, this shows that the functions p_m converge uniformly to f on $[-1,1]$. However, p_m is actually a polynomial. To see this, change the variables and obtain

$$p_m(x) = \int_{x-1}^{x+1} f(t)\,\phi_m(x-t)\,dt$$

which will be a polynomial. To see this, note that a typical term is of the form

$$\int_{x-1}^{x+1} f(t)\,a\,(x-t)^k\,dt,$$

clearly a polynomial in x. This proves Corollary 5.9.2 in case $[a,b]=[-1,1]$. In the general case, there is a linear one to one onto map $l:[-1,1]\to[a,b]$.

$$l(t) = \frac{b-a}{2}(t+1)+a$$

Then if $f\in C([a,b]), f\circ l\in C([-1,1])$. Hence there is a polynomial p such that

$$\max_{t\in[-1,1]} |f\circ l(t) - p(t)| < \varepsilon$$

Then letting $t = l^{-1}(x) = \frac{2(x-a)}{b-a} - 1$, for $x\in[a,b]$,

$$\max_{x\in[a,b]} \left|f(x) - p\left(l^{-1}(x)\right)\right| < \varepsilon$$

but $x\to p\left(l^{-1}(x)\right)$ is a polynomial. This gives an independent proof of that corollary. ■

The next result is the key to the profound generalization of the Weierstrass theorem due to Stone in which an interval will be replaced by a compact and later a locally compact set and polynomials will be replaced with elements of an algebra satisfying certain axioms.

Corollary 5.9.3 *On the interval $[-M,M]$, there exist polynomials p_n such that*

$$p_n(0) = 0$$

and

$$\lim_{n\to\infty} \|p_n - |\cdot|\|_\infty = 0.$$

recall that $\|f\|_\infty \equiv \sup_{t\in[-M,M]} |f(t)|$.

Proof: By Corollary 5.9.2 there exists a sequence of polynomials, $\{\tilde{p}_n\}$ such that $\tilde{p}_n\to|\cdot|$ uniformly. Then let $p_n(t)\equiv\tilde{p}_n(t)-\tilde{p}_n(0)$. ■

Definition 5.9.4 *An algebra of functions, $\mathscr{A}$ defined on A, annihilates no point of A if for all $x\in A$, there exists $g\in\mathscr{A}$ such that $g(x)\neq 0$. The algebra separates points if whenever $x_1\neq x_2$, then there exists $g\in\mathscr{A}$ such that $g(x_1)\neq g(x_2)$.*

The following generalization is known as the Stone Weierstrass approximation theorem.

Theorem 5.9.5 *Let A be a compact topological space and let $\mathscr{A} \subseteq C(A;\mathbb{R})$ be an algebra of functions which separates points and annihilates no point. Then $\mathscr{A}$ is dense in $C(A;\mathbb{R})$.*

Proof: First here is a lemma.

Lemma 5.9.6 *Let c_1 and c_2 be two real numbers and let $x_1 \neq x_2$ be two points of A. Then there exists a function $f_{x_1x_2}$ such that*

$$f_{x_1x_2}(x_1) = c_1,\ f_{x_1x_2}(x_2) = c_2.$$

Proof of the lemma: Let $g \in \mathscr{A}$ satisfy

$$g(x_1) \neq g(x_2).$$

Such a g exists because the algebra separates points. Since the algebra annihilates no point, there exist functions h and k such that

$$h(x_1) \neq 0,\ k(x_2) \neq 0.$$

Then let

$$u \equiv gh - g(x_2)h,\ v \equiv gk - g(x_1)k.$$

It follows that $u(x_1) \neq 0$ and $u(x_2) = 0$ while $v(x_2) \neq 0$ and $v(x_1) = 0$. Let

$$f_{x_1x_2} \equiv \frac{c_1 u}{u(x_1)} + \frac{c_2 v}{v(x_2)}.$$

This proves the lemma. Now continue the proof of Theorem 5.9.5.

First note that $\overline{\mathscr{A}}$ satisfies the same axioms as $\mathscr{A}$ but in addition to these axioms, $\overline{\mathscr{A}}$ is closed. The closure of $\mathscr{A}$ is taken with respect to the usual norm on $C(A)$,

$$\|f\|_\infty \equiv \max\{|f(x)| : x \in A\}.$$

Suppose $f \in \overline{\mathscr{A}}$ and suppose M is large enough that $\|f\|_\infty < M$. Using Corollary 5.9.3, let p_n be a sequence of polynomials such that

$$\|p_n - |\cdot|\|_\infty \to 0,\ p_n(0) = 0.$$

It follows that $p_n \circ f \in \overline{\mathscr{A}}$ and so $|f| \in \overline{\mathscr{A}}$ whenever $f \in \overline{\mathscr{A}}$. Also note that

$$\max(f,g) = \frac{|f-g| + (f+g)}{2}$$

$$\min(f,g) = \frac{(f+g) - |f-g|}{2}.$$

Therefore, this shows that if $f, g \in \overline{\mathscr{A}}$ then $\max(f,g),\ \min(f,g) \in \overline{\mathscr{A}}$. By induction, if $f_i, i = 1,2,\cdots,m$ are in $\overline{\mathscr{A}}$ then

$$\max(f_i, i = 1,2,\cdots,m),\ \min(f_i, i = 1,2,\cdots,m) \in \overline{\mathscr{A}}.$$

Now let $h \in C(A;\mathbb{R})$ and let $x \in A$. Use Lemma 5.9.6 to obtain f_{xy}, a function of $\overline{\mathscr{A}}$ which agrees with h at x and y. Letting $\varepsilon > 0$, there exists an open set $U(y)$ containing y such that

$$f_{xy}(z) > h(z) - \varepsilon \text{ if } z \in U(y).$$

Since A is compact, let $U(y_1), \cdots, U(y_l)$ cover A. Let

$$f_x \equiv \max\left(f_{xy_1}, f_{xy_2}, \cdots, f_{xy_l}\right).$$

Then $f_x \in \overline{\mathscr{A}}$ and $f_x(z) > h(z) - \varepsilon$ for all $z \in A$ and $f_x(x) = h(x)$. This implies that for each $x \in A$ there exists an open set $V(x)$ containing x such that for $z \in V(x)$, $f_x(z) < h(z) + \varepsilon$. Let $V(x_1), \cdots, V(x_m)$ cover A and let $f \equiv \min(f_{x_1}, \cdots, f_{x_m})$. Therefore, $f(z) < h(z) + \varepsilon$ for all $z \in A$ and since $f_x(z) > h(z) - \varepsilon$ for all $z \in A$, it follows $f(z) > h(z) - \varepsilon$ also and so $|f(z) - h(z)| < \varepsilon$ for all z. Since ε is arbitrary, this shows $h \in \overline{\mathscr{A}}$ and proves $\overline{\mathscr{A}} = C(A;\mathbb{R})$. ■

5.10 Connectedness in Normed Linear Space

The main result is that a ball in a normed linear space is connected. This is the next lemma. From this, it follows that for an open set, it is connected if and only if it is arcwise connected.

Lemma 5.10.1 *In a normed vector space, $B(\boldsymbol{z},r)$ is arcwise connected.*

Proof: This is easy from the convexity of the set. If $\boldsymbol{x},\boldsymbol{y} \in B(\boldsymbol{z},r)$, then let $\boldsymbol{\gamma}(t) = \boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})$ for $t \in [0,1]$.

$$\|\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}) - \boldsymbol{z}\| = \|(1-t)(\boldsymbol{x} - \boldsymbol{z}) + t(\boldsymbol{y} - \boldsymbol{z})\|$$

$$\leq (1-t)\|\boldsymbol{x} - \boldsymbol{z}\| + t\|\boldsymbol{y} - \boldsymbol{z}\| < (1-t)r + tr = r$$

showing $\boldsymbol{\gamma}(t)$ stays in $B(\boldsymbol{z},r)$.■

Proposition 5.10.2 *If $X \neq \emptyset$ is arcwise connected, then it is connected.*

Proof: Let $p \in X$. Then by assumption, for any $x \in X$, there is an arc joining p and x. This arc is connected because it is the continuous image of an interval which is connected. Since x is arbitrary, every x is in a connected subset of X which contains p. Hence $C_p = X$ and so X is connected. ■

Theorem 5.10.3 *Let U be an open subset of a normed vector space. Then U is arcwise connected if and only if U is connected. Also the connected components of an open set are open sets.*

Proof: By Proposition 5.10.2 it is only necessary to verify that if U is connected and open in the context of this theorem, then U is arcwise connected. Pick $\boldsymbol{p} \in U$. Say $\boldsymbol{x} \in U$ satisfies $\mathscr{P}$ if there exists a continuous function, $\boldsymbol{\gamma} : [a,b] \to U$ such that $\boldsymbol{\gamma}(a) = \boldsymbol{p}$ and $\boldsymbol{\gamma}(\boldsymbol{b}) = \boldsymbol{x}$.

$$A \equiv \{\boldsymbol{x} \in U \text{ such that } \boldsymbol{x} \text{ satisfies } \mathscr{P}.\}$$

If $\boldsymbol{x} \in A$, then Lemma 5.10.1 implies $B(\boldsymbol{x},r) \subseteq U$ is arcwise connected for small enough r. Thus letting $\boldsymbol{y} \in B(\boldsymbol{x},r)$, there exist intervals, $[a,b]$ and $[c,d]$ and continuous functions

having values in U, γ, η such that $\gamma(a) = p, \gamma(b) = x, \eta(c) = x$, and $\eta(d) = y$. Then let $\gamma_1 : [a, b+d-c] \to U$ be defined as

$$\gamma_1(t) \equiv \begin{cases} \gamma(t) \text{ if } t \in [a,b] \\ \eta(t+c-b) \text{ if } t \in [b, b+d-c] \end{cases}$$

Then it is clear that γ_1 is a continuous function mapping p to y and showing that $B(x,r) \subseteq A$. Therefore, A is open. $A \neq \emptyset$ because since U is open there is an open set, $B(p,\delta)$ containing p which is contained in U and is arcwise connected.

Now consider $B \equiv U \setminus A$. I claim this is also open. If B is not open, there exists a point $z \in B$ such that every open set containing z is not contained in B. Therefore, letting $B(z,\delta)$ be such that $z \in B(z,\delta) \subseteq U$, there exist points of A contained in $B(z,\delta)$. But then, a repeat of the above argument shows $z \in A$ also. Hence B is open and so if $B \neq \emptyset$, then $U = B \cup A$ and so U is separated by the two sets B and A contradicting the assumption that U is connected.

It remains to verify the connected components are open. Let $z \in C_p$ where C_p is the connected component determined by p. Then picking $B(z,\delta) \subseteq U$, $C_p \cup B(z,\delta)$ is connected and contained in U and so it must also be contained in C_p. Thus z is an interior point of C_p. ■

As an application, consider the following corollary.

Corollary 5.10.4 *Let $f : \Omega \to \mathbb{Z}$ be continuous where Ω is a connected open set in a normed vector space. Then f must be a constant.*

Proof: Suppose not. Then it achieves two different values, k and $l \neq k$. Then $\Omega = f^{-1}(l) \cup f^{-1}(\{m \in \mathbb{Z} : m \neq l\})$ and these are disjoint nonempty open sets which separate Ω. To see they are open, note

$$f^{-1}(\{m \in \mathbb{Z} : m \neq l\}) = f^{-1}\left(\cup_{m \neq l}\left(m - \frac{1}{6}, m + \frac{1}{6}\right)\right)$$

which is the inverse image of an open set while $f^{-1}(l) = f^{-1}\left(\left(l - \frac{1}{6}, l + \frac{1}{6}\right)\right)$ also an open set. ■

Definition 5.10.5 *An important concept in a vector space is the concept of convexity. A nonempty set K is called convex if whenever $x, y \in K$, it follows that for all $t \in [0,1]$, $tx + (1-t)y \in K$ also. That is, the line segment joining the two points x, y is in K.*

5.11 Saddle Points*

A very useful idea in nonlinear analysis is the saddle point theorem also called the min max theorem. The proof of this theorem given here follows Brezis [8] which is where I found it. A real valued function f defined on a linear space is convex if

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$$

It is concave if the inequality is turned around. It can be shown that in finite dimensions, convex functions are automatically continuous, similar for concave functions. Recall the following definition of upper and lower semicontinuous functions defined on a metric space and having values in $[-\infty, \infty]$.

Definition 5.11.1 *A function is upper semicontinuous if whenever $x_n \to x$, it follows that $f(x) \geq \limsup_{n\to\infty} f(x_n)$ and it is lower semicontinuous if $f(x) \leq \liminf_{n\to\infty} f(x_n)$.*

The following lemma comes directly from the definition.

Lemma 5.11.2 *If $\mathscr{F}$ is a set of functions which are upper semicontinuous, then $g(x) \equiv \inf\{f(x) : f \in \mathscr{F}\}$ is also upper semicontinuous. Similarly, if $\mathscr{F}$ is a set of functions which are lower semicontinuous, then if $g(x) \equiv \sup\{f(x) : f \in \mathscr{F}\}$ it follows that g is lower semicontinuous.*

Note that in a metric space, the above definitions of upper and lower semicontinuity in terms of sequences are equivalent to the definitions that

$$\begin{aligned} f(x) &\geq \limsup_{r\to 0}\{f(y) : y \in B(x,r)\} \\ f(x) &\leq \liminf_{r\to 0}\{f(y) : y \in B(x,r)\} \end{aligned}$$

respectively.

Here is a technical lemma which will make the proof of the saddle point theorem shorter. It seems fairly interesting also.

Lemma 5.11.3 *Suppose $H : A \times B \to \mathbb{R}$ is strictly convex in the first argument and concave in the second argument where A, B are compact convex nonempty subsets of Banach spaces E, F respectively and $x \to H(x,y)$ is lower semicontinuous while $y \to H(x,y)$ is upper semicontinuous. Let*

$$H(g(y), y) \equiv \min_{x\in A} H(x,y)$$

Then $g(y)$ is uniquely defined and also for $t \in [0,1]$,

$$\lim_{t\to 0} g(y + t(z-y)) = g(y).$$

Proof: First suppose both z, w yield the definition of $g(y)$. Then

$$H\left(\frac{z+w}{2}, y\right) < \frac{1}{2}H(z,y) + \frac{1}{2}H(w,y)$$

which contradicts the definition of $g(y)$. As to the existence of $g(y)$ this is nothing more than the theorem that a lower semicontinuous function defined on a compact set achieves its minimum.

Now consider the last claim about "hemicontinuity", continuity along a line. For all $x \in A$, it follows from the definition of g that

$$H(g(y + t(z-y)), y + t(z-y)) \leq H(x, y + t(z-y))$$

By concavity of H in the second argument,

$$\begin{aligned} & (1-t)H(g(y + t(z-y)), y) + tH(g(y + t(z-y)), z) && (5.7) \\ \leq\ & H(x, y + t(z-y)) && (5.8) \end{aligned}$$

Now let $t_n \to 0$. Does $g(y + t_n(z-y)) \to g(y)$? Suppose not. By compactness, each of $g(y + t_n(z-y))$ is in a compact set and so there is a further subsequence, still denoted by t_n such that

$$g(y + t_n(z-y)) \to \hat{x} \in A$$

Then passing to a limit in 5.8, one obtains, using the upper semicontinuity in one and lower semicontinuity in the other the following inequality.

$$H(\hat{x},y) \leq \lim\inf_{n\to\infty} (1-t_n) H(g(y+t_n(z-y)),y) +$$

$$\lim\inf_{n\to\infty} t_n H(g(y+t_n(z-y)),z)$$

$$\leq \lim\inf_{n\to\infty} \begin{pmatrix} (1-t_n) H(g(y+t_n(z-y)),y) \\ +t_n H(g(y+t_n(z-y)),z) \end{pmatrix}$$

$$\leq \lim\sup_{n\to\infty} H(x,y+t_n(z-y)) \leq H(x,y)$$

This shows that $\hat{x} = g(y)$ because this holds for every x. Since $t_n \to 0$ was arbitrary, this shows that in fact

$$\lim_{t\to 0+} g(y+t(z-y)) = g(y) \ \blacksquare$$

Now with this preparation, here is the min-max theorem.

Definition 5.11.4 *A norm is called strictly convex if whenever $x \neq y$,*

$$\left\| \frac{x+y}{2} \right\| < \frac{\|x\|}{2} + \frac{\|y\|}{2}$$

Theorem 5.11.5 *Let E, F be Banach spaces with E having a strictly convex norm. Also suppose that $A \subseteq E, B \subseteq F$ are compact and convex sets and that $H : A \times B \to \mathbb{R}$ is such that*

$$x \to H(x,y) \text{ is convex}$$

$$y \to H(x,y) \text{ is concave}$$

Assume that $x \to H(x,y)$ is lower semicontinuous and $y \to H(x,y)$ is upper semicontinuous. Then

$$\min_{x\in A} \max_{y\in B} H(x,y) = \max_{y\in B} \min_{x\in A} H(x,y)$$

This condition is equivalent to the existence of $(x_0, y_0) \in A \times B$ such that

$$H(x_0,y) \leq H(x_0,y_0) \leq H(x,y_0) \text{ for all } x,y \tag{5.9}$$

called a saddle point.

Proof: One part of the main equality is obvious.

$$\max_{y\in B} H(x,y) \geq H(x,y) \geq \min_{x\in A} H(x,y)$$

and so for each x,

$$\max_{y\in B} H(x,y) \geq \max_{y\in B} \min_{x\in A} H(x,y)$$

and so

$$\min_{x\in A} \max_{y\in B} H(x,y) \geq \max_{y\in B} \min_{x\in A} H(x,y) \tag{5.10}$$

Next consider the other direction.

Define $H_\varepsilon(x,y) \equiv H(x,y) + \varepsilon \|x\|^2$ where $\varepsilon > 0$. Then H_ε is strictly convex in the first variable. This results from the observation that

$$\left\| \frac{x+y}{2} \right\|^2 < \left(\frac{\|x\| + \|y\|}{2} \right)^2 \leq \frac{1}{2} \left(\|x\|^2 + \|y\|^2 \right),$$

Then by Lemma 5.11.3 there exists a unique $x \equiv g(y)$ such that

$$H_\varepsilon(g(y), y) \equiv \min_{x \in A} H_\varepsilon(x,y)$$

and also, whenever $y, z \in A$,

$$\lim_{t \to 0+} g(y + t(z-y)) = g(y).$$

Thus $H_\varepsilon(g(y), y) = \min_{x \in A} H_\varepsilon(x,y)$. But also this shows that $y \to H_\varepsilon(g(y), y)$ is the minimum of functions which are upper semicontinuous and so this function is also upper semicontinuous. Hence there exists y^* such that

$$\max_{y \in B} H_\varepsilon(g(y), y) = H_\varepsilon(g(y^*), y^*) = \max_{y \in B} \min_{x \in A} H_\varepsilon(x,y) \tag{5.11}$$

Thus from concavity in the second argument and what was just defined, for $t \in (0,1)$,

$$\begin{aligned} H_\varepsilon(g(y^*), y^*) &\geq H_\varepsilon(g((1-t)y^* + ty), (1-t)y^* + ty) \\ &\geq (1-t) H_\varepsilon(g((1-t)y^* + ty), y^*) + t H_\varepsilon(g((1-t)y^* + ty), y) \\ &\geq (1-t) H_\varepsilon(g(y^*), y^*) + t H_\varepsilon(g((1-t)y^* + ty), y) \end{aligned} \tag{5.12}$$

This is because $\min_x H_\varepsilon(x, y^*) \equiv H_\varepsilon(g(y^*), y^*)$ so

$$H_\varepsilon(g((1-t)y^* + ty), y^*) \geq H_\varepsilon(g(y^*), y^*)$$

Then subtracting the first term on the right, one gets

$$t H_\varepsilon(g(y^*), y^*) \geq t H_\varepsilon(g((1-t)y^* + ty), y)$$

and cancelling the t,

$$H_\varepsilon(g(y^*), y^*) \geq H_\varepsilon(g((1-t)y^* + ty), y)$$

Now apply Lemma 5.11.3 and let $t \to 0+$. This along with lower semicontinuity yields

$$H_\varepsilon(g(y^*), y^*) \geq \liminf_{t \to 0+} H_\varepsilon(g((1-t)y^* + ty), y) = H_\varepsilon(g(y^*), y) \tag{5.13}$$

Hence for every x, y

$$H_\varepsilon(x, y^*) \geq H_\varepsilon(g(y^*), y^*) \geq H_\varepsilon(g(y^*), y)$$

Thus

$$\min_x H_\varepsilon(x, y^*) \geq H_\varepsilon(g(y^*), y^*) \geq \max_y H_\varepsilon(g(y^*), y)$$

and so

$$\begin{aligned}\max_{y\in B}\min_{x\in A}H_\varepsilon(x,y) &\geq \min_{x}H_\varepsilon(x,y^*)\geq H_\varepsilon(g(y^*),y^*)\\ &\geq \max_{y}H_\varepsilon(g(y^*),y)\geq \min_{x\in A}\max_{y\in B}H_\varepsilon(x,y)\end{aligned}$$

Thus, letting $C\equiv\max\{\|x\|:x\in A\}$

$$\varepsilon C^2+\max_{y\in B}\min_{x\in A}H(x,y)\geq\min_{x\in A}\max_{y\in B}H(x,y)$$

Since ε is arbitrary, it follows that

$$\max_{y\in B}\min_{x\in A}H(x,y)\geq\min_{x\in A}\max_{y\in B}H(x,y)$$

This proves the first part because it was shown above in 5.10 that

$$\min_{x\in A}\max_{y\in B}H(x,y)\geq\max_{y\in B}\min_{x\in A}H(x,y)$$

Now consider 5.9 about the existence of a "saddle point" given the equality of min max and max min. Let

$$\alpha=\max_{y\in B}\min_{x\in A}H(x,y)=\min_{x\in A}\max_{y\in B}H(x,y)$$

Then from

$$y\to\min_{x\in A}H(x,y)\text{ and }x\to\max_{y\in B}H(x,y)$$

being upper semicontinuous and lower semicontinuous respectively, there exist y_0 and x_0 such that

$$\alpha=\min_{x\in A}H(x,y_0)=\max_{y\in B}\overset{\text{minimum of u.s.c}}{\overset{\cdot}{\min_{x\in A}}}H(x,y)=\min_{x\in A}\overset{\text{maximum of l.s.c.}}{\max_{y\in B}}H(x,y)=\max_{y\in B}H(x_0,y)$$

Then

$$\begin{aligned}\alpha &= \max_{y\in B}H(x_0,y)\geq H(x_0,y_0)\\ \alpha &= \min_{x\in A}H(x,y_0)\leq H(x_0,y_0)\end{aligned}$$

so in fact $\alpha=H(x_0,y_0)$ and from the above equalities,

$$\begin{aligned}H(x_0,y_0) &= \alpha=\min_{x\in A}H(x,y_0)\leq H(x,y_0)\\ H(x_0,y_0) &= \alpha=\max_{y\in B}H(x_0,y)\geq H(x_0,y)\end{aligned}$$

and so

$$H(x_0,y)\leq H(x_0,y_0)\leq H(x,y_0)$$

Thus if the min max condition holds, then there exists a saddle point, namely (x_0,y_0).

Finally suppose there is a saddle point (x_0,y_0) where

$$H(x_0,y)\leq H(x_0,y_0)\leq H(x,y_0)$$

Then

$$\min_{x\in A}\max_{y\in B} H(x,y) \leq \max_{y\in B} H(x_0,y) \leq H(x_0,y_0) \leq \min_{x\in A} H(x,y_0) \leq \max_{y\in B}\min_{x\in A} H(x,y)$$

However, as noted above, it is always the case that

$$\max_{y\in B}\min_{x\in A} H(x,y) \leq \min_{x\in A}\max_{y\in B} H(x,y) \quad \blacksquare$$

What was really needed? You needed compactness of A, B and these sets needed to be in a linear space. Of course there needed to be a norm for which $x \to \|x\|$ is strictly convex and lower semicontinuous, so the conditions given above are sufficient but maybe not necessary. You might try generalizing this much later after reading about weak topologies.

5.12 Exercises

1. Consider the metric space $C([0,T],\mathbb{R}^n)$ with the norm

$$\|\boldsymbol{f}\| \equiv \max_{x\in[0,T]} \|\boldsymbol{f}(x)\|_\infty$$

Explain why the maximum exists. Show this is a complete metric space. **Hint:** If you have $\{\boldsymbol{f}_m\}$ a Cauchy sequence in $C([0,T],\mathbb{R}^n)$, then for each x, you have $\{\boldsymbol{f}_m(x)\}$ a Cauchy sequence in $\mathbb{R}^n$. Recall that this is a complete space. Thus there exists $\boldsymbol{f}(x) = \lim_{m\to\infty} \boldsymbol{f}_m(x)$. You must show that $\boldsymbol{f}$ is continuous. This was in the section on the Ascoli Arzela theorem in more generality if you need an outline of how this goes. Write down the details for this case. Note how $\boldsymbol{f}$ is in bold face. This means it is a function which has values in $\mathbb{R}^n$. $\boldsymbol{f}(t) = (f_1(t), f_2(t), \cdots, f_n(t))$.

2. For $\boldsymbol{f} \in C([0,T],\mathbb{R}^n)$, you define the Riemann integral in the usual way using Riemann sums. Alternatively, you can define it as

$$\int_0^t \boldsymbol{f}(s)\,ds = \left(\int_0^t f_1(s)\,ds, \int_0^t f_2(s)\,ds, \cdots, \int_0^t f_n(s)\,ds\right)$$

Then show that the following limit exists in $\mathbb{R}^n$ for each $t \in (0,T)$.

$$\lim_{h\to 0} \frac{\int_0^{t+h} \boldsymbol{f}(s)\,ds - \int_0^t \boldsymbol{f}(s)\,ds}{h} = \boldsymbol{f}(t).$$

You should use the fundamental theorem of calculus from one variable calculus and the definition of the norm to verify this. As a review, in case we don't get to it in time, for $\boldsymbol{f}$ defined on an interval $[0,T]$ and $s \in [0,T]$, $\lim_{t\to s} \boldsymbol{f}(t) = \boldsymbol{l}$ means that for all $\varepsilon > 0$, there exists $\delta > 0$ such that if $0 < |t-s| < \delta$, then $\|\boldsymbol{f}(t) - \boldsymbol{l}\|_\infty < \varepsilon$.

3. Suppose $f : \mathbb{R} \to \mathbb{R}$ and $f \geq 0$ on $[-1,1]$ with $f(-1) = f(1) = 0$ and $f(x) < 0$ for all $x \notin [-1,1]$. Can you use a modification of the proof of the Weierstrass approximation theorem for functions on an interval presented earlier to show that for all $\varepsilon > 0$ there exists a polynomial p, such that $|p(x) - f(x)| < \varepsilon$ for $x \in [-1,1]$ and $p(x) \leq 0$ for all $x \notin [-1,1]$? **Hint:**Let $f_\varepsilon(x) = f(x) - \frac{\varepsilon}{2}$. Thus there exists δ such that $1 > \delta > 0$ and $f_\varepsilon < 0$ on $(-1,-1+\delta)$ and $(1-\delta,1)$. Now consider $\phi_k(x) = a_k\left(\delta^2 - \left(\frac{x}{\delta}\right)^2\right)^k$ and try something similar to the proof given for the Weierstrass approximation theorem above.

4. A collection of functions $\mathscr{F}$ of $C([0,T],\mathbb{R}^n)$ is said to be uniformly equicontinuous if for every $\varepsilon > 0$ there exists $\delta > 0$ such that if $\boldsymbol{f} \in \mathscr{F}$ and $|t-s| < \delta$, then $\|\boldsymbol{f}(t) - \boldsymbol{f}(s)\|_\infty < \varepsilon$. Thus the functions are uniformly continuous all at once. The single δ works for every pair t,s closer together than δ and for all functions $\boldsymbol{f} \in \mathscr{F}$. As an easy case, suppose there exists K such that for all $\boldsymbol{f} \in \mathscr{F}$, $\|\boldsymbol{f}(t) - \boldsymbol{f}(s)\|_\infty \leq K|t-s|$. Show that $\mathscr{F}$ is uniformly equicontinuous. Now suppose $\mathscr{G}$ is a collection of functions of $C([0,T],\mathbb{R}^n)$ which is bounded. That is, $\|\boldsymbol{f}\| = \max_{t\in[0,T]} \|\boldsymbol{f}(t)\|_\infty < M < \infty$ for all $\boldsymbol{f} \in \mathscr{G}$. Then let $\mathscr{F}$ denote the functions which are of the form $\boldsymbol{F}(t) \equiv \boldsymbol{y}_0 + \int_0^t \boldsymbol{f}(s)\,ds$ where $\boldsymbol{f} \in \mathscr{G}$. Show that $\mathscr{F}$ is uniformly equicontinuous. **Hint:** This is a really easy problem if you do the right things. Here is the way you should proceed. Remember the triangle inequality from one variable calculus which said that for $a < b$ $\left|\int_a^b f(s)\,ds\right| \leq \int_a^b |f(s)|\,ds$. Then $\left\|\int_a^b \boldsymbol{f}(s)\,ds\right\|_\infty = \max_i \left|\int_a^b f_i(s)\,ds\right| \leq \max_i \int_a^b |f_i(s)|\,ds \leq \int_a^b \|\boldsymbol{f}(s)\|_\infty\,ds$. Reduce to the case just considered using the assumption that these $\boldsymbol{f}$ are bounded.

5. Suppose $\mathscr{F}$ is a set of functions in $C([0,T],\mathbb{R}^n)$ which is uniformly bounded and uniformly equicontinuous as described above. Show it must be totally bounded. **Hint:** Let $\varepsilon > 0$ be given. Then let $0 = t_0 < t_1 < \cdots < t_p = T$ where this is a uniform partition of the interval $[0,T]$ and each $t_j - t_{j-1} = \delta/2$ where δ is such that if $|s-t| < \delta$, then $\|\boldsymbol{f}(t) - \boldsymbol{f}(s)\|_\infty < \varepsilon/8$ for all $\boldsymbol{f} \in \mathscr{F}$. If $\mathscr{F}$ is not totally bounded, then there exists ε and δ as just described, along with a sequence $\{\boldsymbol{f}_k\}$ where $\|\boldsymbol{f}_k - \boldsymbol{f}_m\| \geq \varepsilon$ for all $k \neq m$. Take a subsequence p times corresponding to the partition just described to obtain the existence of a subsequence $\left\{\boldsymbol{f}_{k_m}\right\}_{m=1}^\infty$ such that $\left\{\boldsymbol{f}_{k_m}(t_j)\right\}$ is a Cauchy sequence for each of these t_j in the partition. Call it $\{\boldsymbol{f}_m\}$ to save notation. Now exploit uniform equicontinuity to obtain a contradiction. It goes like this. Pick t. Then it is within δ of some t_j. Then

$$\begin{aligned}\|\boldsymbol{f}_m(t) - \boldsymbol{f}_k(t)\|_\infty \quad &\leq \quad \left\|\boldsymbol{f}_m(t) - \boldsymbol{f}_m(t_j)\right\|_\infty + \left\|\boldsymbol{f}_m(t_j) - \boldsymbol{f}_k(t_j)\right\|_\infty \\ &\quad + \left\|\boldsymbol{f}_k(t_j) - \boldsymbol{f}_k(t)\right\|_\infty\end{aligned}$$

   Now note that since there are finitely many t_j, there is a single N such that $k,m > N$ implies the middle term is less than $\varepsilon/8$ for all t_j at once. Get a contradiction from the use of equicontinuity to contradict the assumption that $\|\boldsymbol{f}_k - \boldsymbol{f}_m\| \geq \varepsilon$ for all $k \neq m$.

6. ↑If $A \subseteq (X,d)$ is totally bounded, show that $\bar{A}$ the closure of A is also totally bounded. In the above problem, explain why $\bar{\mathscr{F}}$ the closure of $\mathscr{F}$ is compact. This uses the big theorem on compactness. Try and do this on your own, but if you get stuck, it is in the section on Arzela Ascoli theorem. When you have done this problem, you have proved the important part of the Arzela Ascoli theorem in the special case where the functions are defined on an interval. You can use this to prove one of the most important results in the theory of differential equations. This theorem is a really profound result because it gives compactness in a normed linear space which is **not finite dimensional.** Thus this is a non trivial generalization of the Heine Borel theorem.

7. Let $(X,\|\cdot\|)$ be a normed linear space. A set A is said to be **convex** if whenever $\boldsymbol{x},\boldsymbol{y} \in A$ the line segment determined by these points given by $t\boldsymbol{x} + (1-t)\boldsymbol{y}$ for $t \in [0,1]$ is

also in A. Show that every open or closed ball is convex. Remember a closed ball is $D(\boldsymbol{x},r) \equiv \{\hat{\boldsymbol{x}} : \|\hat{\boldsymbol{x}}-\boldsymbol{x}\| \leq r\}$ while the open ball is $B(\boldsymbol{x},r) \equiv \{\hat{\boldsymbol{x}} : \|\hat{\boldsymbol{x}}-\boldsymbol{x}\| < r\}$. This should work just as easily in any normed linear space with any norm.

8. Let K be a nonempty closed and convex set in an inner product space $(X,|\cdot|)$ which is complete. For example, $\mathbb{F}^n$ or any other finite dimensional inner product space. Let $y \notin K$ and let $\lambda = \inf\{|y-x| : x \in K\}$. Let $\{x_n\}$ be a minimizing sequence. That is $\lambda = \lim_{n\to\infty} |y-x_n|$. Explain why such a minimizing sequence exists. Next explain the following using the parallelogram identity in the above problem as follows.

$$\left|y - \frac{x_n+x_m}{2}\right|^2 = \left|\frac{y}{2} - \frac{x_n}{2} + \frac{y}{2} - \frac{x_m}{2}\right|^2$$

$$= -\left|\frac{y}{2} - \frac{x_n}{2} - \left(\frac{y}{2} - \frac{x_m}{2}\right)\right|^2 + \frac{1}{2}|y-x_n|^2 + \frac{1}{2}|y-x_m|^2$$

Hence

$$\begin{aligned}\left|\frac{x_m-x_n}{2}\right|^2 &= -\left|y - \frac{x_n+x_m}{2}\right|^2 + \frac{1}{2}|y-x_n|^2 + \frac{1}{2}|y-x_m|^2 \\ &\leq -\lambda^2 + \frac{1}{2}|y-x_n|^2 + \frac{1}{2}|y-x_m|^2\end{aligned}$$

Next explain why the right hand side converges to 0 as $m,n \to \infty$. Thus $\{x_n\}$ is a Cauchy sequence and converges to some $x \in X$. Explain why $x \in K$ and $|x-y| = \lambda$. Thus there exists a closest point in K to y. Next show that there is only one closest point. **Hint:** To do this, suppose there are two x_1, x_2 and consider $\frac{x_1+x_2}{2}$ using the parallelogram law to show that this average works better than either of the two points which is a contradiction unless they are really the same point. This theorem is of enormous significance.

9. Let K be a closed convex nonempty set in a complete inner product space $(H,|\cdot|)$ (Hilbert space) and let $y \in H$. Denote the closest point to y by Py. Show that Py is characterized as being the solution to the following variational inequality

$$\mathrm{Re}\,(z-Py, y-Py) \leq 0$$

for all $z \in K$. That is, show that $x = Py$ if and only if $\mathrm{Re}\,(z-x, y-x) \leq 0$ for all $z \in K$. **Hint:** Let $x \in K$. Then, due to convexity, a generic thing in K is of the form $x + t(z-x), t \in [0,1]$ for every $z \in K$. Then

$$|x+t(z-x)-y|^2 = |x-y|^2 + t^2|z-x|^2 - t2\mathrm{Re}\,(z-x, y-x)$$

If $x = Px$, then the minimum value of this on the left occurs when $t = 0$. Function defined on $[0,1]$ has its minimum at $t = 0$. What does it say about the derivative of this function at $t = 0$? Next consider the case that for some x the inequality $\mathrm{Re}\,(z-x, y-x) \leq 0$. Explain why this shows $x = Py$.

10. Using Problem 9 and Problem 8 show the projection map, P onto a closed convex subset is Lipschitz continuous with Lipschitz constant 1. That is $|Px-Py| \leq |x-y|$.

11. Suppose, in an inner product space, you know $\mathrm{Re}(x,y)$. Show that you also know $\mathrm{Im}(x,y)$. That is, give a formula for $\mathrm{Im}(x,y)$ in terms of $\mathrm{Re}(x,y)$. **Hint:**

$$\begin{aligned}(x,iy) &= -i(x,y) = -i(\mathrm{Re}(x,y)+i\mathrm{Im}(x,y)) \\ &= -i\mathrm{Re}(x,y)+\mathrm{Im}(x,y)\end{aligned}$$

while, by definition, $(x,iy)=\mathrm{Re}(x,iy)+i\mathrm{Im}(x,iy)$. Now consider matching real and imaginary parts.

12. Let $h>0$ be given and let $\boldsymbol{f}(t,\boldsymbol{x})\in\mathbb{R}^n$ for each $\boldsymbol{x}\in\mathbb{R}^n$. Also let $(t,\boldsymbol{x})\to\boldsymbol{f}(t,\boldsymbol{x})$ be continuous and $\sup_{t,\boldsymbol{x}}\|\boldsymbol{f}(t,\boldsymbol{x})\|_\infty<C<\infty$. Let $\boldsymbol{x}_h(t)$ be a solution to the following

$$\boldsymbol{x}_h(t)=\boldsymbol{x}_0+\int_0^t \boldsymbol{f}(s,\boldsymbol{x}_h(s-h))\,ds$$

where $\boldsymbol{x}_h(s-h)\equiv\boldsymbol{x}_0$ if $s-h\le 0$. Explain why there exists a solution. **Hint:** Consider the intervals $[0,h],[h,2h]$ and so forth. Next explain why these functions $\{\boldsymbol{x}_h\}_{h>0}$ are equicontinuous and uniformly bounded. Now use the result of Problem 6 to argue that there exists a subsequence, still denoted by $\boldsymbol{x}_h$ such that $\lim_{h\to 0}\boldsymbol{x}_h=\boldsymbol{x}$ in $C([0,T];\mathbb{R}^n)$ as discussed in Problem 5. Use what you learned about the Riemann integral in single variable advanced calculus to explain why you can pass to a limit and conclude that $\boldsymbol{x}(t)=\boldsymbol{x}_0+\int_0^t\boldsymbol{f}(s,\boldsymbol{x}(s))\,ds$ **Hint:**

$$\begin{aligned}&\left\|\int_0^t \boldsymbol{f}(s,\boldsymbol{x}(s))\,ds-\int_0^t \boldsymbol{f}(s,\boldsymbol{x}_h(s-h))\,ds\right\|_\infty \\ \le\ &\left\|\int_0^t \boldsymbol{f}(s,\boldsymbol{x}(s))\,ds-\int_0^t \boldsymbol{f}(s,\boldsymbol{x}(s-h))\,ds\right\|_\infty \\ &+\left\|\int_0^t \boldsymbol{f}(s,\boldsymbol{x}(s-h))\,ds-\int_0^t \boldsymbol{f}(s,\boldsymbol{x}_h(s-h))\,ds\right\|_\infty \\ \le\ &\int_0^T\|\boldsymbol{f}(s,\boldsymbol{x}(s))-\boldsymbol{f}(s,\boldsymbol{x}(s-h))\|_\infty\,ds \\ &+\int_0^T\|\boldsymbol{f}(s,\boldsymbol{x}(s-h))-\boldsymbol{f}(s,\boldsymbol{x}_h(s-h))\|_\infty\,ds\end{aligned}$$

Now use Problem 2 to verify that $\boldsymbol{x}'=\boldsymbol{f}(t,\boldsymbol{x})$, $\boldsymbol{x}(0)=\boldsymbol{x}_0$. When you have done this, you will have proved the celebrated Peano existence theorem from ordinary differential equations.

13. Let $|\alpha|\equiv\sum_i\alpha_i$. Let $\mathscr{G}$ denote all finite sums of functions of the form $p(\boldsymbol{x})e^{-a|\boldsymbol{x}|^2}$ where $p(\boldsymbol{x})$ is a polynomial and $a>0$. If you consider all real valued continuous functions defined on the closed ball $\overline{B(\mathbf{0},R)}$ show that if f is such a function, then for every $\varepsilon>0$, there exists $g\in\mathscr{G}$ such that $\|f-g\|_\infty<\varepsilon$ where $\|h\|_\infty\equiv\max_{\boldsymbol{x}\in\overline{B(\mathbf{0},R)}}|h(\boldsymbol{x})|$. Thus, from multi-variable calculus, every continuous function f is uniformly close to an infinitely differentiable function on any closed ball centered at $\mathbf{0}$.

14. Suppose now that $f \in C_0(\mathbb{R}^p)$. This means that f is everywhere continuous and that $\lim_{\|\boldsymbol{x}\|\to\infty} |f(\boldsymbol{x})| = 0$. Show that for every $\varepsilon > 0$ there exists $g \in \mathscr{G}$ such that $\sup_{\boldsymbol{x}\in\mathbb{R}^p} |f(\boldsymbol{x}) - g(\boldsymbol{x})| < \varepsilon$. Thus you can approximate such a continuous function f uniformly on all of $\mathbb{R}^p$ with a function which has infinitely many continuous partial derivatives. I assume the reader has had a beginning course in multi-variable calculus including partial derivatives. If not, a partial derivative is just a derivative with respect to one of the variables, fixing all the others.

15. In Problem 23 on Page 124, and $V \equiv \operatorname{span}(f_{p_1}, ..., f_{p_n}), f_r(x) \equiv x^r, x \in [0,1]$ and $-\frac{1}{2} < p_1 < p_2 < \cdots$ with $\lim_{k\to\infty} p_k = \infty$. The distance between f_m and V is

$$\frac{1}{\sqrt{2m+1}} \prod_{j\le n} \frac{|m-p_j|}{(p_j+m+1)} = d$$

Let $d_n = d$ so more functions are allowed to be included in V. Show that $\sum_n \frac{1}{p_n} = \infty$ if and only if $\lim_{n\to\infty} d_n = 0$. Explain, using the Weierstrass approximation theorem why this shows that if g is a function continuous on $[0,1]$, then there is a function $\sum_{k=1}^N a_k f_{p_k}$ with $\left|g - \sum_{k=1}^N a_k f_{p_k}\right| < \varepsilon$. Here $|g|^2 \equiv \int_0^1 |g(x)|^2 dx$. This is Müntz's first theorem. **Hint:** $d_n \to 0$, if and only if $\ln d_n \to -\infty$ so you might want to arrange things so that this happens. You might want to use the fact that for $x \in [0,1/2], -x \ge \ln(1-x) \ge -2x$. See [10] which is where I read this. That product is $\prod_{j\le n}\left(1 - \left(1 - \frac{|m-p_j|}{(p_j+m+1)}\right)\right)$ and so ln of this expression is

$$\sum_{j=1}^n \ln\left(1 - \left(1 - \frac{|m-p_j|}{(p_j+m+1)}\right)\right)$$

which is in the interval

$$\left[-2\sum_{j=1}^n \left(1 - \frac{|m-p_j|}{(p_j+m+1)}\right), -\sum_{j=1}^n \left(1 - \frac{|m-p_j|}{(p_j+m+1)}\right)\right]$$

and so $d_n \to 0$ if and only if $\sum_{j=1}^\infty \left(1 - \frac{|m-p_j|}{(p_j+m+1)}\right) = \infty$. Since $p_n \to \infty$ it suffices to consider the convergence of $\sum_j \left(1 - \frac{p_j-m}{(p_j+m+1)}\right) = \sum_j \left(\frac{2m+1}{(p_j+m+1)}\right)$. Now recall theorems from calculus.

16. For $f \in C([a,b];\mathbb{R})$, real valued continuous functions, let

$$|f| \equiv \left(\int_a^b |f(t)|^2\right)^{1/2} \equiv (f,f)^{1/2}$$

where $(f,g) \equiv \int_a^b f(x)\, g(x)\, dx$. Recall the Cauchy Schwarz inequality

$$|(f,g)| \le |f|\,|g|$$

Now suppose $\frac{1}{2} < p_1 < p_2 \cdots$ where $\lim_{k\to\infty} p_k = \infty$. Let

$$V_n = \operatorname{span}(1, f_{p_1}, f_{p_2}, ..., f_{p_n}).$$

For $\|\cdot\|$ the uniform approximation norm, show that for every $g \in C([0,1])$, there exists there exists a sequence of functions, $f_n \in V_n$ such that $\|g - f_n\| \to 0$. This is the second Müntz theorem. **Hint:** Show that you can approximate $x \to x^m$ uniformly. To do this, use the above Müntz to approximate mx^{m-1} with $\sum_k c_k x^{p_k - 1}$ in the inner product norm.

$$\int_0^1 \left| mx^{m-1} - \sum_{k=1}^n c_k x^{p_k-1} \right|^2 dx \leq \varepsilon^2$$

Then $x^m - \sum_{k=1}^n \frac{c_k}{p_k} x^{p_k} = \int_0^x \left(mt^{m-1} - \sum_{k=1}^n c_k t^{p_k-1} \right) dt$. Then

$$\begin{aligned} \left| x^m - \sum_{k=1}^n \frac{c_k}{p_k} x^{p_k} \right| &\leq \int_0^x \left| mt^{m-1} - \sum_{k=1}^n c_k t^{p_k-1} \right| dt \\ &\leq \int_0^1 1 \left| mt^{m-1} - \sum_{k=1}^n c_k t^{p_k-1} \right| dt. \end{aligned}$$

Now use the Cauchy Schwarz inequality on that last integral to obtain

$$\max_{x\in[0,1]} \left| x^m - \sum_{k=1}^n \frac{c_k}{p_k} x^{p_k} \right| \leq \varepsilon.$$

In case $m = 0$, there is nothing to show because 1 is in V_n. Explain why the result follows from this and the Weierstrass approximation theorem.

Chapter 6

Fixed Point Theorems

This is on fixed point theorems which feature the Brouwer fixed point theorem. This next block of material is a discussion of simplices and triangulations used to prove the Brouwer fixed point theorem in an elementary way. It features the famous Sperner's lemma and is based on very elementary concepts from linear algebra in an essential way. However, it is pretty technical stuff. This elementary proof is harder than those which come from other approaches like integration theory or degree theory. These other shorter ways of obtaining the Brouwer fixed point theorem from analytical methods are presented later. If desired, this chapter could be placed after the easier to prove version of the Brouwer fixed point theorem, Theorem 11.11.9 on Page 344 after sufficient integration theory has been presented. I like the approach presented in this chapter which is based on simplices because it is elementary and contains a method for locating a fixed point. It seems philosophically wrong to make this theorem depend on integration theory.

6.1 Simplices and Triangulations

Definition 6.1.1 *Define an n simplex, denoted by* $[\boldsymbol{x}_0,\cdots,\boldsymbol{x}_n]$*, to be the convex hull of the* $n+1$ *points,* $\{\boldsymbol{x}_0,\cdots,\boldsymbol{x}_n\}$ *where* $\{\boldsymbol{x}_i-\boldsymbol{x}_0\}_{i=1}^n$ *are linearly independent. Thus*

$$[\boldsymbol{x}_0,\cdots,\boldsymbol{x}_n]\equiv\left\{\sum_{i=0}^n t_i\boldsymbol{x}_i:\sum_{i=0}^n t_i=1,\ t_i\geq 0\right\}.$$

Note that $\{\boldsymbol{x}_j-\boldsymbol{x}_m\}_{j\neq m}$ *are also independent. I will call the* $\{t_i\}$ *just described the coordinates of a point* $\boldsymbol{x}$*.*

To see the last claim, suppose $\sum_{j\neq m}c_j(\boldsymbol{x}_j-\boldsymbol{x}_m)=0$. Then you would have

$$c_0(\boldsymbol{x}_0-\boldsymbol{x}_m)+\sum_{j\neq m,0}c_j(\boldsymbol{x}_j-\boldsymbol{x}_m)=0$$

$$=c_0(\boldsymbol{x}_0-\boldsymbol{x}_m)+\sum_{j\neq m,0}c_j(\boldsymbol{x}_j-\boldsymbol{x}_0)+\left(\sum_{j\neq m,0}c_j\right)(\boldsymbol{x}_0-\boldsymbol{x}_m)=0$$

$$=\sum_{j\neq m,0}c_j(\boldsymbol{x}_j-\boldsymbol{x}_0)+\left(\sum_{j\neq m}c_j\right)(\boldsymbol{x}_0-\boldsymbol{x}_m)$$

Then you get $\sum_{j\neq m}c_j=0$ and each $c_j=0$ for $j\neq m,0$. Thus $c_0=0$ also because the sum is 0 and all other $c_j=0$.

Since $\{\boldsymbol{x}_i-\boldsymbol{x}_0\}_{i=1}^n$ is an independent set, the t_i used to specify a point in the convex hull are uniquely determined. If two of them are $\sum_{i=0}^n t_i\boldsymbol{x}_i=\sum_{i=0}^n s_i\boldsymbol{x}_i$.Then $\sum_{i=0}^n t_i(\boldsymbol{x}_i-\boldsymbol{x}_0)=\sum_{i=0}^n s_i(\boldsymbol{x}_i-\boldsymbol{x}_0)$ so $t_i=s_i$ for $i\geq 1$ by independence. Since the s_i and t_i sum to 1, it follows that also $s_0=t_0$. If $n\leq 2$, the simplex is a triangle, line segment, or point. If $n\leq 3$, it is a tetrahedron, triangle, line segment or point.

Definition 6.1.2 *If S is an n simplex. Then it is triangulated if it is the union of smller sub-simplices, the triangulation, such that if* S_1,S_2 *are two simplices in the triangulation, with*

$$S_1\equiv\left[\boldsymbol{z}_0^1,\cdots,\boldsymbol{z}_m^1\right],\ S_2\equiv\left[\boldsymbol{z}_0^2,\cdots,\boldsymbol{z}_p^2\right]$$

then

$$S_1 \cap S_2 = \left[\boldsymbol{x}_{k_0}, \cdots, \boldsymbol{x}_{k_r}\right]$$

where $\left[\boldsymbol{x}_{k_0}, \cdots, \boldsymbol{x}_{k_r}\right]$ *is in the triangulation and*

$$\left\{\boldsymbol{x}_{k_0}, \cdots, \boldsymbol{x}_{k_r}\right\} = \left\{\boldsymbol{z}_0^1, \cdots, \boldsymbol{z}_m^1\right\} \cap \left\{\boldsymbol{z}_0^2, \cdots, \boldsymbol{z}_p^2\right\}$$

or else the two simplices do not intersect.

The following proposition is geometrically fairly clear. It will be used without comment whenever needed in the following argument about triangulations.

Proposition 6.1.3 *Say* $[\boldsymbol{x}_1, \cdots, \boldsymbol{x}_r], [\hat{\boldsymbol{x}}_1, \cdots, \hat{\boldsymbol{x}}_r], [\boldsymbol{z}_1, \cdots, \boldsymbol{z}_r]$ *are all* $r-1$ *simplices and*

$$[\boldsymbol{x}_1, \cdots, \boldsymbol{x}_r], [\hat{\boldsymbol{x}}_1, \cdots, \hat{\boldsymbol{x}}_r] \subseteq [\boldsymbol{z}_1, \cdots, \boldsymbol{z}_r]$$

and $[\boldsymbol{z}_1, \cdots, \boldsymbol{z}_r, \boldsymbol{b}]$ *is an* $r+1$ *simplex and*

$$[\boldsymbol{y}_1, \cdots, \boldsymbol{y}_s] = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_r] \cap [\hat{\boldsymbol{x}}_1, \cdots, \hat{\boldsymbol{x}}_r] \tag{6.1}$$

where

$$\{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_s\} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_r\} \cap \{\hat{\boldsymbol{x}}_1, \cdots, \hat{\boldsymbol{x}}_r\} \tag{6.2}$$

Then

$$[\boldsymbol{x}_1, \cdots, \boldsymbol{x}_r, \boldsymbol{b}] \cap [\hat{\boldsymbol{x}}_1, \cdots, \hat{\boldsymbol{x}}_r, \boldsymbol{b}] = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_s, \boldsymbol{b}] \tag{6.3}$$

Proof: If you have $\sum_{i=1}^s t_i \boldsymbol{y}_i + t_{s+1}\boldsymbol{b}$ in the right side, the t_i summing to 1 and nonnegative, then it is obviously in both of the two simplices on the left because of 6.2. Thus $[\boldsymbol{x}_1, \cdots, \boldsymbol{x}_r, \boldsymbol{b}] \cap [\hat{\boldsymbol{x}}_1, \cdots, \hat{\boldsymbol{x}}_r, \boldsymbol{b}] \supseteq [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_s, \boldsymbol{b}]$.

Now suppose $\boldsymbol{x}_k = \sum_{j=1}^r t_j^k \boldsymbol{z}_j, \hat{\boldsymbol{x}}_k = \sum_{j=1}^r \hat{t}_j^k \boldsymbol{z}_j$, as usual, the scalars adding to 1 and nonnegative.

Consider something in both of the simplices on the left in 6.3. Is it in the right? The element on the left is of the form

$$\sum_{\alpha=1}^r s_\alpha \boldsymbol{x}_\alpha + s_{r+1}\boldsymbol{b} = \sum_{\alpha=1}^r \hat{s}_\alpha \hat{\boldsymbol{x}}_\alpha + \hat{s}_{r+1}\boldsymbol{b}$$

where the s_α, are nonnegative and sum to one, similarly for $\hat{s}_\alpha$. Thus

$$\sum_{j=1}^r \sum_{\alpha=1}^r s_\alpha t_j^\alpha \boldsymbol{z}_j + s_{r+1}\boldsymbol{b} = \sum_{\alpha=1}^r \sum_{j=1}^r \hat{s}_\alpha \hat{t}_j^\alpha \boldsymbol{z}_j + \hat{s}_{r+1}\boldsymbol{b} \tag{6.4}$$

Now observe that

$$\sum_j \sum_\alpha s_\alpha t_j^\alpha + s_{r+1} = \sum_\alpha \sum_j s_\alpha t_j^\alpha + s_{r+1} = \sum_\alpha s_\alpha + s_{r+1} = 1.$$

A similar observation holds for the right side of 6.4. By uniqueness of the coordinates in an $r+1$ simplex, and assumption that $[\boldsymbol{z}_1, \cdots, \boldsymbol{z}_r, \boldsymbol{b}]$ is an $r+1$ simplex, $\hat{s}_{r+1} = s_{r+1}$ and so

$$\sum_{\alpha=1}^r \frac{s_\alpha}{1-s_{r+1}} \boldsymbol{x}_\alpha = \sum_{\alpha=1}^r \frac{\hat{s}_\alpha}{1-s_{r+1}} \hat{\boldsymbol{x}}_\alpha$$

where $\sum_\alpha \frac{s_\alpha}{1-s_{r+1}} = \sum_\alpha \frac{\hat{s}_\alpha}{1-s_{r+1}} = 1$, which would say that both sides are a single element of $[\boldsymbol{x}_1, \cdots, \boldsymbol{x}_r] \cap [\hat{\boldsymbol{x}}_1, \cdots, \hat{\boldsymbol{x}}_r] = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_s]$ and this shows both are equal to something of the form $\sum_{i=1}^s t_i \boldsymbol{y}_i, \sum_i t_i = 1, t_i \geq 0$. Therefore,

$$\sum_{\alpha=1}^r \frac{s_\alpha}{1-s_{r+1}} \boldsymbol{x}_\alpha = \sum_{i=1}^s t_i \boldsymbol{y}_i, \sum_{\alpha=1}^r s_\alpha \boldsymbol{x}_\alpha = \sum_{i=1}^s (1-s_{r+1}) t_i \boldsymbol{y}_i$$

It follows that

$$\sum_{\alpha=1}^r s_\alpha \boldsymbol{x}_\alpha + s_{r+1}\boldsymbol{b} = \sum_{i=1}^s (1-s_{r+1}) t_i \boldsymbol{y}_i + s_{r+1}\boldsymbol{b} \in [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_s, \boldsymbol{b}]$$

which proves the other inclusion. ■

Next I will explain why any simplex can be triangulated in such a way that all subsimplices have diameter less than ε.

This is obvious if $n \leq 2$. Supposing it to be true for $n-1$, is it also so for n? The barycenter $\boldsymbol{b}$ of a simplex $[\boldsymbol{x}_0, \cdots, \boldsymbol{x}_n]$ is $\frac{1}{1+n}\sum_i \boldsymbol{x}_i$. This point is not in the convex hull of any of the faces, those simplices of the form $[\boldsymbol{x}_0, \cdots, \hat{\boldsymbol{x}}_k, \cdots, \boldsymbol{x}_n]$ where the hat indicates $\boldsymbol{x}_k$ has been left out. Thus, placing $\boldsymbol{b}$ in the k^{th} position, $[\boldsymbol{x}_0, \cdots, \boldsymbol{b}, \cdots, \boldsymbol{x}_n]$ is a n simplex also. First note that $[\boldsymbol{x}_0, \cdots, \hat{\boldsymbol{x}}_k, \cdots, \boldsymbol{x}_n]$ is an $n-1$ simplex. To be sure $[\boldsymbol{x}_0, \cdots, \boldsymbol{b}, \cdots, \boldsymbol{x}_n]$ is an n simplex, we need to check that certain vectors are linearly independent. If

$$\boldsymbol{0} = \sum_{j=1}^{k-1} c_j (\boldsymbol{x}_j - \boldsymbol{x}_0) + a_k \left(\frac{1}{n+1} \sum_{i=0}^n \boldsymbol{x}_i - \boldsymbol{x}_0 \right) + \sum_{j=k+1}^n d_j (\boldsymbol{x}_j - \boldsymbol{x}_0)$$

then does it follow that $a_k = 0 = c_j = d_j$?

$$\boldsymbol{0} = \sum_{j=1}^{k-1} c_j (\boldsymbol{x}_j - \boldsymbol{x}_0) + a_k \frac{1}{n+1} \left(\sum_{i=0}^n (\boldsymbol{x}_i - \boldsymbol{x}_0) \right) + \sum_{j=k+1}^n d_j (\boldsymbol{x}_j - \boldsymbol{x}_0)$$

$$\begin{aligned} \boldsymbol{0} &= \sum_{j=1}^{k-1} \left(c_j + \frac{a_k}{n+1} \right) (\boldsymbol{x}_j - \boldsymbol{x}_0) + a_k \frac{1}{n+1} (\boldsymbol{x}_k - \boldsymbol{x}_0) \\ &\quad + \sum_{j=k+1}^n \left(d_j + \frac{a_k}{n+1} \right) (\boldsymbol{x}_j - \boldsymbol{x}_0) \end{aligned}$$

Thus $\frac{a_k}{n+1} = 0$ and each $c_j + \frac{a_k}{n+1} = 0 = d_j + \frac{a_k}{n+1}$ so each c_j and d_j are also 0. Thus, this is also an n simplex.

Actually, a little more is needed. Suppose $[\boldsymbol{y}_0, \cdots, \boldsymbol{y}_{n-1}]$ is an $n-1$ simplex such that $[\boldsymbol{y}_0, \cdots, \boldsymbol{y}_{n-1}] \subseteq [\boldsymbol{x}_0, \cdots, \hat{\boldsymbol{x}}_k, \cdots, \boldsymbol{x}_n]$. Why is $[\boldsymbol{y}_0, \cdots, \boldsymbol{y}_{n-1}, \boldsymbol{b}]$ an n simplex? We know the vectors $\{\boldsymbol{y}_j - \boldsymbol{y}_0\}_{k=1}^{n-1}$ are independent and that $\boldsymbol{y}_j = \sum_{i \neq k} t_i^j \boldsymbol{x}_i$ where $\sum_{i \neq k} t_i^j = 1$ with each being nonnegative. Suppose

$$\sum_{j=1}^{n-1} c_j (\boldsymbol{y}_j - \boldsymbol{y}_0) + c_n (\boldsymbol{b} - \boldsymbol{y}_0) = \boldsymbol{0} \tag{6.5}$$

If $c_n = 0$, then by assumption, each $c_j = 0$. The proof goes by assuming $c_n \neq 0$ and deriving a contradiction. Assume then that $c_n \neq 0$. Then you can divide by it and obtain modified

constants, still denoted as c_j such that

$$\boldsymbol{b} = \frac{1}{n+1}\sum_{i=0}^{n}\boldsymbol{x}_i = \boldsymbol{y}_0 + \sum_{j=1}^{n-1} c_j\left(\boldsymbol{y}_j - \boldsymbol{y}_0\right)$$

Thus

$$\frac{1}{n+1}\sum_{i=0}^{n}\sum_{s\neq k} t_s^0\left(\boldsymbol{x}_i - \boldsymbol{x}_s\right) = \sum_{j=1}^{n-1} c_j\left(\boldsymbol{y}_j - \boldsymbol{y}_0\right) = \sum_{j=1}^{n-1} c_j\left(\sum_{s\neq k} t_s^j \boldsymbol{x}_s - \sum_{s\neq k} t_s^0 \boldsymbol{x}_s\right)$$

$$= \sum_{j=1}^{n-1} c_j\left(\sum_{s\neq k} t_s^j\left(\boldsymbol{x}_s - \boldsymbol{x}_0\right) - \sum_{s\neq k} t_s^0\left(\boldsymbol{x}_s - \boldsymbol{x}_0\right)\right)$$

Modify the term on the left and simplify on the right to get

$$\frac{1}{n+1}\sum_{i=0}^{n}\sum_{s\neq k} t_s^0\left(\left(\boldsymbol{x}_i - \boldsymbol{x}_0\right) + \left(\boldsymbol{x}_0 - \boldsymbol{x}_s\right)\right) = \sum_{j=1}^{n-1} c_j\left(\sum_{s\neq k}\left(t_s^j - t_s^0\right)\left(\boldsymbol{x}_s - \boldsymbol{x}_0\right)\right)$$

Thus,

$$\begin{aligned}\frac{1}{n+1}\sum_{i=0}^{n}\left(\sum_{s\neq k} t_s^0\right)\left(\boldsymbol{x}_i - \boldsymbol{x}_0\right) &= \frac{1}{n+1}\sum_{i=0}^{n}\sum_{s\neq k} t_s^0\left(\boldsymbol{x}_s - \boldsymbol{x}_0\right) \\ &\quad + \sum_{j=1}^{n-1} c_j\left(\sum_{s\neq k}\left(t_s^j - t_s^0\right)\left(\boldsymbol{x}_s - \boldsymbol{x}_0\right)\right)\end{aligned}$$

Then, taking out the $i = k$ term on the left yields

$$\frac{1}{n+1}\left(\sum_{s\neq k} t_s^0\right)\left(\boldsymbol{x}_k - \boldsymbol{x}_0\right) = -\frac{1}{n+1}\sum_{i\neq k}\left(\sum_{s\neq k} t_s^0\right)\left(\boldsymbol{x}_i - \boldsymbol{x}_0\right)$$

$$\frac{1}{n+1}\sum_{i=0}^{n}\sum_{s\neq k} t_s^0\left(\boldsymbol{x}_s - \boldsymbol{x}_0\right) + \sum_{j=1}^{n-1} c_j\left(\sum_{s\neq k}\left(t_s^j - t_s^0\right)\left(\boldsymbol{x}_s - \boldsymbol{x}_0\right)\right)$$

That on the right is a linear combination of vectors $\left(\boldsymbol{x}_r - \boldsymbol{x}_0\right)$ for $r \neq k$ so by independence, $\sum_{r\neq k} t_r^0 = 0$. However, each $t_r^0 \geq 0$ and these sum to 1 so this is impossible. Hence $c_n = 0$ after all and so each $c_j = 0$. Thus $[\boldsymbol{y}_0, \cdots, \boldsymbol{y}_{n-1}, \boldsymbol{b}]$ is an n simplex.

Now in general, if you have an n simplex $[\boldsymbol{x}_0, \cdots, \boldsymbol{x}_n]$, its diameter is the maximum of $|\boldsymbol{x}_k - \boldsymbol{x}_l|$ for all $k \neq l$. Consider $\left|\boldsymbol{b} - \boldsymbol{x}_j\right|$. It equals

$$\left|\sum_{i=0}^{n}\frac{1}{n+1}\left(\boldsymbol{x}_i - \boldsymbol{x}_j\right)\right| = \left|\sum_{i\neq j}\frac{1}{n+1}\left(\boldsymbol{x}_i - \boldsymbol{x}_j\right)\right| \leq \frac{n}{n+1}\operatorname{diam}(S).$$

Consider the k^{th} face of S which is the simplex $[\boldsymbol{x}_0, \cdots, \hat{\boldsymbol{x}}_k, \cdots, \boldsymbol{x}_n]$. By induction, it has a triangulation into simplices which each have diameter no more than $\frac{n}{n+1}\operatorname{diam}(S)$. Let these $n-1$ simplices be denoted by $\left\{S_1^k, \cdots, S_{m_k}^k\right\}$. Then the simplices $\left\{\left[S_i^k, \boldsymbol{b}\right]\right\}_{i=1,k=1}^{m_k,n+1}$ are a triangulation of S such that $\operatorname{diam}\left(\left[S_i^k, \boldsymbol{b}\right]\right) \leq \frac{n}{n+1}\operatorname{diam}(S)$. Do for $\left[S_i^k, \boldsymbol{b}\right]$ what was just done for S obtaining a triangulation of S as the union of what is obtained such that each simplex has diameter no more than $\left(\frac{n}{n+1}\right)^2\operatorname{diam}(S)$. Continuing this way shows the existence of the desired triangulation. You simply do the process k times where $\left(\frac{n}{n+1}\right)^k\operatorname{diam}(S) < \varepsilon$.

6.2 Labeling Vertices

Next is a way to label the vertices. Let $p_0, \cdots, p_n$ be the first $n+1$ prime numbers. All vertices of a simplex $S = [\boldsymbol{x}_0, \cdots, \boldsymbol{x}_n]$ having $\{\boldsymbol{x}_k - \boldsymbol{x}_0\}_{k=1}^n$ independent will be labeled with one of these primes. In particular, the vertex $\boldsymbol{x}_k$ will be labeled as p_k if the simplex is $[\boldsymbol{x}_0, \cdots, \boldsymbol{x}_n]$. The "value" of a simplex will be the product of its labels. Triangulate this S.

Consider a 1 simplex whose vertices are from the vertices of S, the original n simplex $\left[\boldsymbol{x}_{k_1}, \boldsymbol{x}_{k_2}\right]$, label $\boldsymbol{x}_{k_1}$ as p_{k_1} and $\boldsymbol{x}_{k_2}$ as p_{k_2}. Then label all other vertices of this triangulation which occur on $\left[\boldsymbol{x}_{k_1}, \boldsymbol{x}_{k_2}\right]$ either p_{k_1} or p_{k_2}. Note that by independence of $\{\boldsymbol{x}_k - \boldsymbol{x}_r\}_{k \neq r}$, this cannot introduce an inconsistency because the segment cannot contain any other vertex of S. Then obviously there will be an odd number of simplices in this triangulation having value $p_{k_1} p_{k_2}$, that is a p_{k_1} at one end and a p_{k_2} at the other. Next consider the 2 simplices $\left[\boldsymbol{x}_{k_1}, \boldsymbol{x}_{k_2}, \boldsymbol{x}_{k_3}\right]$ where the $\boldsymbol{x}_{k_i}$ are from S. Label all vertices of the triangulation which lie on one of these 2 simplices which have not already been labeled as either p_{k_1}, p_{k_2}, or p_{k_2}. Continue this way. This labels all vertices of the triangulation of S which have at least one coordinate zero. For the vertices of the triangulation which have all coordinates positive, the interior points of S, label these at random from any of $p_0, ..., p_n$. (Essentially, this is the same idea. The "interior" points are the new ones not already labeled.) The idea is to show that there is an odd number of n simplices with value $\prod_{i=0}^n p_i$ in the triangulation and more generally, for each m simplex $\left[\boldsymbol{x}_{k_1}, \cdots, \boldsymbol{x}_{k_{m+1}}\right], m \leq n$ with the $\boldsymbol{x}_{k_i}$ an original vertex from S, there are an odd number of m simplices of the triangulation contained in $\left[\boldsymbol{x}_{k_1}, \cdots, \boldsymbol{x}_{k_{m+1}}\right]$, having value $p_{k_1} \cdots p_{k_{m+1}}$. It is clear that this is the case for all such 1 simplices. For convenience, call such simplices $\left[\boldsymbol{x}_{k_1}, \cdots, \boldsymbol{x}_{k_{m+1}}\right]$ m dimensional faces of S. An m simplex which is a subspace of this one will have the "correct" value if its value is $p_{k_1} \cdots p_{k_{m+1}}$.

Suppose that the labeling has produced an odd number of simplices of the triangulation contained in each m dimensional face of S which have the correct value. Take such an m dimensional face $\left[\boldsymbol{x}_{j_1}, \ldots, \boldsymbol{x}_{j_{k+1}}\right]$. Consider $\hat{S} \equiv$

$$\left[\boldsymbol{x}_{j_1}, \ldots \boldsymbol{x}_{j_{k+1}}, \boldsymbol{x}_{j_{k+2}}\right]$$

Then by induction, there is an odd number of k simplices on the s^{th} face

$$\left[\boldsymbol{x}_{j_1}, \ldots, \hat{\boldsymbol{x}}_{j_s}, \cdots, \boldsymbol{x}_{j_{k+2}}\right]$$

having value $\prod_{i \neq s} p_{j_i}$. In particular, the face $\left[\boldsymbol{x}_{j_1}, \ldots, \boldsymbol{x}_{j_{k+1}}, \hat{\boldsymbol{x}}_{j_{k+2}}\right]$ has an odd number of simplices with value $\prod_{i \leq k+1} p_{j_i}$.

No simplex in any other face of $\hat{S}$ can have this value by uniqueness of prime factorization. Pick a simplex on the face $\left[\boldsymbol{x}_{j_1}, \ldots, \boldsymbol{x}_{j_{k+1}}, \hat{\boldsymbol{x}}_{j_{k+2}}\right]$ which has correct value $\prod_{i \leq k+1} p_{j_i}$ and cross this simplex into $\hat{S}$. Continue crossing simplices having value $\prod_{i \leq k+1} p_{j_i}$ which have not been crossed till the process ends. It must end because there are an odd number of these simplices having value $\prod_{i \leq k+1} p_{j_i}$. If the process leads to the outside of $\hat{S}$, then one can always enter it again because there are an odd number of simplices with value $\prod_{i \leq k+1} p_{j_i}$ available and you will have used up an even number. Note that in this process, if you have a simplex with one side labeled $\prod_{i \leq k+1} p_{j_i}$, there is either one way in or out of this simplex or two depending on whether the remaining vertex is labeled $p_{j_{k+2}}$. When the process ends, the value of the simplex must be $\prod_{i=1}^{k+2} p_{j_i}$ because it will have the additional label $p_{j_{k+2}}$. Otherwise, there would be another route out of this, through the other side labeled $\prod_{i \leq k+1} p_{j_i}$. This identifies a simplex in the triangulation with value $\prod_{i=1}^{k+2} p_{j_i}$.

Then repeat the process with $\prod_{i\leq k+1} p_{j_i}$ valued simplices on $\left[\boldsymbol{x}_{j_1},\ldots,\boldsymbol{x}_{j_{k+1}},\hat{\boldsymbol{x}}_{j_{k+2}}\right]$ which have not been crossed. Repeating the process, entering from the outside, cannot deliver a $\prod_{i=1}^{k+2} p_{j_i}$ valued simplex encountered earlier because of what was just noted. There is either one or two ways to cross the simplices. In other words, the process is one to one in selecting a $\prod_{i\leq k+1} p_{j_i}$ simplex from crossing such a simplex on the selected face of $\hat{S}$. Continue doing this, crossing a $\prod_{i\leq k+1} p_{j_i}$ simplex on the face of $\hat{S}$ which has not been crossed previously. This identifies an odd number of simplices having value $\prod_{i=1}^{k+2} p_{j_i}$. These are the ones which are "accessible" from the outside using this process. If there are any which are not accessible from outside, applying the same process starting inside one of these, leads to exactly one other inaccessible simplex with value $\prod_{i=1}^{k+2} p_{j_i}$. Hence these inaccessible simplices occur in pairs and so there are an odd number of simplices in the triangulation having value $\prod_{i=1}^{k+2} p_{j_i}$. We refer to this procedure of labeling as Sperner's lemma. The system of labeling is well defined thanks to the assumption that $\{\boldsymbol{x}_k-\boldsymbol{x}_0\}_{k=1}^n$ is independent which implies that $\{\boldsymbol{x}_k-\boldsymbol{x}_i\}_{k\neq i}$ is also linearly independent. Thus there can be no ambiguity in the labeling of vertices on any "face" the convex hull of some of the original vertices of S. The following is a description of the system of labeling the vertices.

Lemma 6.2.1 *Let $[\boldsymbol{x}_0,\cdots,\boldsymbol{x}_n]$ be an n simplex with $\{\boldsymbol{x}_k-\boldsymbol{x}_0\}_{k=1}^n$ independent, and let the first $n+1$ primes be $p_0,p_1,\cdots,p_n$. Label $\boldsymbol{x}_k$ as p_k and consider a triangulation of this simplex. Labeling the vertices of this triangulation which occur on $\left[\boldsymbol{x}_{k_1},\cdots,\boldsymbol{x}_{k_s}\right]$ with any of $p_{k_1},\cdots,p_{k_s}$, beginning with all 1 simplices $\left[\boldsymbol{x}_{k_1},\boldsymbol{x}_{k_2}\right]$ and then 2 simplices and so forth, there are an odd number of simplices $\left[\boldsymbol{y}_{k_1},\cdots,\boldsymbol{y}_{k_s}\right]$ of the triangulation contained in $\left[\boldsymbol{x}_{k_1},\cdots,\boldsymbol{x}_{k_s}\right]$ which have value $p_{k_1}\cdots p_{k_s}$. This for $s=1,2,\cdots,n$.*

A combinatorial method

We now give a brief discussion of the system of labeling for Sperner's lemma from the point of view of counting numbers of faces rather than obtaining them with an algorithm. Let $p_0,\cdots,p_n$ be the first $n+1$ prime numbers. All vertices of a simplex $S=[\boldsymbol{x}_0,\cdots,\boldsymbol{x}_n]$ having $\{\boldsymbol{x}_k-\boldsymbol{x}_0\}_{k=1}^n$ independent will be labeled with one of these primes. In particular, the vertex $\boldsymbol{x}_k$ will be labeled as p_k. The value of a simplex will be the product of its labels. Triangulate this S. Consider a 1 simplex coming from the original simplex $\left[\boldsymbol{x}_{k_1},\boldsymbol{x}_{k_2}\right]$, label one end as p_{k_1} and the other as p_{k_2}. Then label all other vertices of this triangulation which occur on $\left[\boldsymbol{x}_{k_1},\boldsymbol{x}_{k_2}\right]$ either p_{k_1} or p_{k_2}. The assumption of linear independence assures that no **other** vertex of S can be in $\left[\boldsymbol{x}_{k_1},\boldsymbol{x}_{k_2}\right]$ so there will be no inconsistency in the labeling. Then obviously there will be an odd number of simplices in this triangulation having value $p_{k_1}p_{k_2}$, that is a p_{k_1} at one end and a p_{k_2} at the other. Suppose that the labeling has been done for all vertices of the triangulation which are on $\left[\boldsymbol{x}_{j_1},\ldots\boldsymbol{x}_{j_{k+1}}\right]$,

$$\left\{\boldsymbol{x}_{j_1},\ldots\boldsymbol{x}_{j_{k+1}}\right\}\subseteq\{\boldsymbol{x}_0,\ldots\boldsymbol{x}_n\}$$

any k simplex for $k\leq n-1$, and there is an odd number of simplices from the triangulation having value equal to $\prod_{i=1}^{k+1} p_{j_i}$. Consider $\hat{S}\equiv\left[\boldsymbol{x}_{j_1},\ldots\boldsymbol{x}_{j_{k+1}},\boldsymbol{x}_{j_{k+2}}\right]$. Then by induction, there is an odd number of k simplices on the s^{th} face

$$\left[\boldsymbol{x}_{j_1},\ldots,\hat{\boldsymbol{x}}_{j_s},\cdots,\boldsymbol{x}_{j_{k+1}}\right]$$

having value $\prod_{i\neq s} p_{j_i}$. In particular the face $\left[\boldsymbol{x}_{j_1},\ldots,\boldsymbol{x}_{j_{k+1}},\hat{\boldsymbol{x}}_{j_{k+2}}\right]$ has an odd number of simplices with value $\prod_{i=1}^{k+1} p_{j_i}:=\hat{P}_k$. We want to argue that some simplex in the triangulation which is contained in $\hat{S}$ has value $\hat{P}_{k+1}:=\prod_{i=1}^{k+2} p_{j_i}$. Let Q be the number of $k+1$

simplices from the triangulation contained in $\hat{S}$ which have two faces with value $\hat{P}_k$ (A $k+1$ simplex has either 1 or 2 $\hat{P}_k$ faces.) and let R be the number of $k+1$ simplices from the triangulation contained in $\hat{S}$ which have exactly one $\hat{P}_k$ face. These are the ones we want because they have value $\hat{P}_{k+1}$. Thus the number of faces having value $\hat{P}_k$ which is described here is $2Q+R$. All interior $\hat{P}_k$ faces being counted twice by this number. Now we count the total number of $\hat{P}_k$ faces another way. There are P of them on the face $\left[\boldsymbol{x}_{j_1},\ldots,\boldsymbol{x}_{j_{k+1}},\hat{\boldsymbol{x}}_{j_{k+2}}\right]$ and by induction, P is odd. Then there are O of them which are not on this face. These faces got counted twice. Therefore,

$$2Q+R=P+2O$$

and so, since P is odd, so is R. Thus there is an odd number of $\hat{P}_{k+1}$ simplices in $\hat{S}$.

We refer to this procedure of labeling as Sperner's lemma. The system of labeling is well defined thanks to the assumption that $\{\boldsymbol{x}_k-\boldsymbol{x}_0\}_{k=1}^n$ is independent which implies that $\{\boldsymbol{x}_k-\boldsymbol{x}_i\}_{k\neq i}$ is also linearly independent. Thus there can be no ambiguity in the labeling of vertices on any "face", the convex hull of some of the original vertices of S. Sperner's lemma is now a consequence of this discussion.

6.3 The Brouwer Fixed Point Theorem

$S\equiv[\boldsymbol{x}_0,\cdots,\boldsymbol{x}_n]$ is a simplex in $\mathbb{R}^n$. Assume $\{\boldsymbol{x}_i-\boldsymbol{x}_0\}_{i=1}^n$ are linearly independent. Thus a typical point of S is of the form $\sum_{i=0}^n t_i\boldsymbol{x}_i$ where the t_i are uniquely determined and the map $\boldsymbol{x}\to\boldsymbol{t}$ is continuous from S to the compact set

$$\left\{\boldsymbol{t}\in\mathbb{R}^{n+1}:\sum t_i=1,t_i\geq 0\right\}$$

The map $\boldsymbol{t}\to\boldsymbol{x}$ is one to one and clearly continuous. Since S is compact, it follows that the inverse map is also continuous. This is a general consideration but what follows is a short explanation why this is so in this specific example.

To see this, suppose $\boldsymbol{x}^k\to\boldsymbol{x}$ in S. Let $\boldsymbol{x}^k\equiv\sum_{i=0}^n t_i^k\boldsymbol{x}_i$ with $\boldsymbol{x}$ defined similarly with t_i^k replaced with t_i, $\boldsymbol{x}\equiv\sum_{i=0}^n t_i\boldsymbol{x}_i$. Then

$$\boldsymbol{x}^k-\boldsymbol{x}_0=\sum_{i=0}^n t_i^k\boldsymbol{x}_i-\sum_{i=0}^n t_i^k\boldsymbol{x}_0=\sum_{i=1}^n t_i^k\left(\boldsymbol{x}_i-\boldsymbol{x}_0\right)$$

Thus

$$\boldsymbol{x}^k-\boldsymbol{x}_0=\sum_{i=1}^n t_i^k\left(\boldsymbol{x}_i-\boldsymbol{x}_0\right),\ \boldsymbol{x}-\boldsymbol{x}_0=\sum_{i=1}^n t_i\left(\boldsymbol{x}_i-\boldsymbol{x}_0\right)$$

Say t_i^k fails to converge to t_i for all $i\geq 1$. Then there exists a subsequence, still denoted with superscript k such that for each $i=1,\cdots,n$, it follows that $t_i^k\to s_i$ where $s_i\geq 0$ and some $s_i\neq t_i$. But then, taking a limit, it follows that

$$\boldsymbol{x}-\boldsymbol{x}_0=\sum_{i=1}^n s_i\left(\boldsymbol{x}_i-\boldsymbol{x}_0\right)=\sum_{i=1}^n t_i\left(\boldsymbol{x}_i-\boldsymbol{x}_0\right)$$

which contradicts independence of the $\boldsymbol{x}_i-\boldsymbol{x}_0$. It follows that for all $i\geq 1, t_i^k\to t_i$. Since they all sum to 1, this implies that also $t_0^k\to t_0$. Thus the claim about continuity is verified.

Let $\boldsymbol{f}:S\to S$ be continuous. When doing $\boldsymbol{f}$ to a point $\boldsymbol{x}$, one obtains another point of S denoted as $\sum_{i=0}^n s_i\boldsymbol{x}_i$. Thus in this argument the scalars s_i will be the components after doing $\boldsymbol{f}$ to a point of S denoted as $\sum_{i=0}^n t_i\boldsymbol{x}_i$.

Consider a triangulation of S such that all simplices in the triangulation have diameter less than ε. The vertices of the simplices in this triangulation will be labeled from $p_0, \cdots, p_n$ the first $n+1$ prime numbers. If $[\boldsymbol{y}_0, \cdots, \boldsymbol{y}_n]$ is one of these simplices in the triangulation, each vertex is of the form $\sum_{l=0}^n t_l \boldsymbol{x}_l$ where $t_l \geq 0$ and $\sum_l t_l = 1$. Let $\boldsymbol{y}_i$ be one of these vertices, $\boldsymbol{y}_i = \sum_{l=0}^n t_l \boldsymbol{x}_l$, the t_l being determined by $\boldsymbol{y}_i$. Define $r_j \equiv s_j/t_j$ if $t_j > 0$ and ∞ if $t_j = 0$. Then $p(\boldsymbol{y}_i)$ will be the label placed on $\boldsymbol{y}_i$. To determine this label, let r_k be the smallest of these ratios. Then the label placed on $\boldsymbol{y}_i$ will be p_k where r_k is the smallest of all these extended nonnegative real numbers just described. If there is duplication, pick p_k where k is smallest. The value of the simplex will be the product of the labels. What does it mean for the value of the simplex to be $P_n \equiv p_0 p_1 \cdots p_n$? It means that each of the first $n+1$ primes is assigned to exactly one of the $n+1$ vertices of the simplex so each $r_j > 0$ and there are no repeats in the r_j.

Note that for the vertices which are on $[\boldsymbol{x}_{i_1}, \cdots, \boldsymbol{x}_{i_m}]$, these will be labeled from the list $\{p_{i_1}, \cdots, p_{i_m}\}$ because $t_k = 0$ for each of these and so $r_k = \infty$ unless $k \in \{i_1, \cdots, i_m\}$. In particular, this scheme labels $\boldsymbol{x}_i$ as p_i.

By the Sperner's lemma procedure described above, there are an odd number of simplices having value $\prod_{i \neq k} p_i$ on the k^{th} face and an odd number of simplices in the triangulation of S for which the value of the simplex is $p_0 p_1 \cdots p_n \equiv P_n$. Thus if $[\boldsymbol{y}_0, \cdots, \boldsymbol{y}_n]$ is one of these simplices, and $p(\boldsymbol{y}_i)$ is the label for $\boldsymbol{y}_i$, $\prod_{i=0}^n p(\boldsymbol{y}_i) = \prod_{j=0}^n p_j \equiv P_n$

What is r_k, the smallest of those ratios in determining a label? Could it be larger than 1? r_k is certainly finite because at least some $t_j \neq 0$ since they sum to 1. Thus, if $r_k > 1$, you would have $s_k > t_k$. The s_j sum to 1 and so some $s_j < t_j$ since otherwise, the sum of the t_j equalling 1 would require the sum of the s_j to be larger than 1. Hence r_k was not really the smallest after all and so $r_k \leq 1$. Hence $s_k \leq t_k$. Thus if the value of a simplex is P_n, then for each vertex of the simplex, the smallest ratio associated with it is of the form $s_j/t_j \leq 1$ and each j gets used exactly once.

Let $\mathscr{S} \equiv \{S_1, \cdots, S_m\}$ denote those simplices whose value is P_n. In other words, if $\{\boldsymbol{y}_0, \cdots, \boldsymbol{y}_n\}$ are the vertices of one of these simplices in $\mathscr{S}$, and $\boldsymbol{y}_s = \sum_{i=0}^n t_i^s \boldsymbol{x}_i$, $r_{k_s} \leq r_j$ for all $j \neq k_s$ and $\{k_0, \cdots, k_n\} = \{0, \cdots, n\}$. Let $\boldsymbol{b}$ denote the barycenter of $S_k = [\boldsymbol{y}_0, \cdots, \boldsymbol{y}_n]$. $\boldsymbol{b} \equiv \sum_{i=0}^n \frac{1}{n+1} \boldsymbol{y}_i$

Do the same system of labeling for each n simplex in a sequence of triangulations where the diameters of the simplices in the k^{th} triangulation are no more than 2^{-k}. Thus each of these triangulations has a n simplex having diameter no more than 2^{-k} which has value P_n. Let $\boldsymbol{b}_k$ be the barycenter of one of these n simplices having value P_n. By compactness, there is a subsequence, still denoted with the index k such that $\boldsymbol{b}_k \to \boldsymbol{x}$. This $\boldsymbol{x}$ is a fixed point.

Consider this last claim. $\boldsymbol{x} = \sum_{i=0}^n t_i \boldsymbol{x}_i$ and after applying $\boldsymbol{f}$, the result is $\sum_{i=0}^n s_i \boldsymbol{x}_i$. Then $\boldsymbol{b}_k$ is the barycenter of some σ_k having diameter no more than 2^{-k} which has value P_n. Say σ_k is a simplex having vertices $\left\{\boldsymbol{y}_0^k, \cdots, \boldsymbol{y}_n^k\right\}$ and the value of $\left[\boldsymbol{y}_0^k, \cdots, \boldsymbol{y}_n^k\right]$ is P_n. Thus also $\lim_{k \to \infty} \boldsymbol{y}_i^k = \boldsymbol{x}$. Re ordering these vertices if necessary, we can assume that the label for $\boldsymbol{y}_i^k$ is p_i which implies that the smallest ratio r_k is when $k = i$ and as noted above, this ratio is no larger than 1. Thus for each $i = 0, \cdots, n$,

$$\frac{s_i}{t_i} \leq 1, \; s_i \leq t_i$$

the i^{th} coordinate of $\boldsymbol{f}\left(\boldsymbol{y}_i^k\right)$ with respect to the original vertices of S decreases and each i is represented for $i = \{0, 1, \cdots, n\}$. As noted above, $\boldsymbol{y}_i^k \to \boldsymbol{x}$ and so the i^{th} coordinate of $\boldsymbol{y}_i^k, t_i^k$ must converge to t_i. Hence if the i^{th} coordinate of $\boldsymbol{f}\left(\boldsymbol{y}_i^k\right)$ is denoted by s_i^k, $s_i^k \leq t_i^k$. By continuity of $\boldsymbol{f}$, it follows that $s_i^k \to s_i$. Thus the above inequality is preserved on taking

$k \to \infty$ and so $0 \leq s_i \leq t_i$ this for each i and these s_i, t_i pertain to the single point $\boldsymbol{x}$. But these s_i add to 1 as do the t_i and so in fact, $s_i = t_i$ for each i and so $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{x}$. This proves the following theorem which is the Brouwer fixed point theorem.

Theorem 6.3.1 *Let S be a simplex $[\boldsymbol{x}_0, \cdots, \boldsymbol{x}_n]$ such that $\{\boldsymbol{x}_i - \boldsymbol{x}_0\}_{i=1}^n$ are independent. Also let $\boldsymbol{f} : S \to S$ be continuous. Then there exists $\boldsymbol{x} \in S$ such that $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{x}$.*

Corollary 6.3.2 *Let K be a closed convex bounded subset of $\mathbb{R}^n$. Let $\boldsymbol{f} : K \to K$ be continuous. Then there exists $\boldsymbol{x} \in K$ such that $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{x}$.*

Proof: Let S be a large simplex containing K and let P be the projection map onto K. See Problem 10 on Page 152 for the necessary properties of this projection map. Consider $\boldsymbol{g}(\boldsymbol{x}) \equiv \boldsymbol{f}(P\boldsymbol{x})$. Then $\boldsymbol{g}$ satisfies the necessary conditions for Theorem 6.3.1 and so there exists $\boldsymbol{x} \in S$ such that $\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{x}$. But this says $\boldsymbol{x} \in K$ and so $\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x})$. ■

Definition 6.3.3 *A set B has the fixed point property if whenever $f : B \to B$ for f continuous, it follows that f has a fixed point.*

The proof of this corollary is pretty significant. By a homework problem, a closed convex set is a retract of $\mathbb{R}^n$. This is what it means when you say there is this continuous projection map which maps onto the closed convex set but does not change any point in the closed convex set. When you have a set A which is a subset of a set B which has the property that continuous functions $\boldsymbol{f} : B \to B$ have fixed points, and there is a continuous map P from B to A which leaves points of A unchanged, then it follows that A will have the same "fixed point property". You can probably imagine all sorts of sets which are retracts of closed convex bounded sets. Also, if you have a compact set B which has the fixed point property and $h : B \to h(B)$ with h one to one and continuous, it will follow that h^{-1} is continuous and that $h(B)$ will also have the fixed point property. This is very easy to show. This will allow further extensions of this theorem. This says that the fixed point property is topological.

Several of the following theorems are generalizations of the Brouwer fixed point theorem.

6.4 The Schauder Fixed Point Theorem

First we give a proof of the Schauder fixed point theorem which is an infinite dimensional generalization of the Brouwer fixed point theorem. This is a theorem which lives in Banach space. Recal that one of these is a complete normed vector space. There is also a version of this theorem valid in locally convex topological vector spaces where the theorem is sometimes called Tychonoff's theorem. In infinite dimensions, the closed unit ball fails to have the fixed point property. Thus something more is needed to get a fixed point.

We let K be a closed convex subset of X a Banach space and let

$$f \text{ be continuous, } f : K \to K, \text{ and } \overline{f(K)} \text{ is compact.}$$

Lemma 6.4.1 *For each $r > 0$ there exists a finite set of points*

$$\{y_1, \cdots, y_n\} \subseteq \overline{f(K)}$$

and continuous functions ψ_i *defined on* $\overline{f(K)}$ *such that for* $x \in \overline{f(K)}$,

$$\sum_{i=1}^{n} \psi_i(x) = 1, \tag{6.6}$$

$$\psi_i(x) = 0 \text{ if } x \notin B(y_i, r),\ \psi_i(x) > 0 \text{ if } x \in B(y_i, r).$$

If

$$f_r(x) \equiv \sum_{i=1}^{n} y_i \psi_i(f(x)), \tag{6.7}$$

then whenever $x \in K$,

$$\|f(x) - f_r(x)\| \leq r.$$

Proof: Using the compactness of $\overline{f(K)}$ which implies this set is totally bounded, there exists an r net

$$\{y_1, \cdots, y_n\} \subseteq \overline{f(K)} \subseteq K$$

such that $\{B(y_i, r)\}_{i=1}^{n}$ covers $\overline{f(K)}$. Let

$$\phi_i(y) \equiv (r - \|y - y_i\|)^+$$

Thus $\phi_i(y) > 0$ if $y \in B(y_i, r)$ and $\phi_i(y) = 0$ if $y \notin B(y_i, r)$. For $x \in \overline{f(K)}$, let

$$\psi_i(x) \equiv \phi_i(x) \left(\sum_{j=1}^{n} \phi_j(x) \right)^{-1}.$$

Then 6.6 is satisfied. Indeed the denominator is not zero because x is in one of the $B(y_i, r)$. Thus it is obvious that the sum of these $\psi_i(f(x))$ equals 1 for $x \in K$. Now let f_r be given by 6.7 for $x \in K$. For such x,

$$f(x) - f_r(x) = \sum_{i=1}^{n} (f(x) - y_i) \psi_i(f(x))$$

Thus

$$\begin{aligned} f(x) - f_r(x) &= \sum_{\{i: f(x) \in B(y_i, r)\}} (f(x) - y_i) \psi_i(f(x)) \\ &\quad + \sum_{\{i: f(x) \notin B(y_i, r)\}} (f(x) - y_i) \psi_i(f(x)) \\ &= \sum_{\{i: f(x) - y_i \in B(0, r)\}} (f(x) - y_i) \psi_i(f(x)) = \\ &\sum_{\{i: f(x) - y_i \in B(0, r)\}} (f(x) - y_i) \psi_i(f(x)) + \sum_{\{i: f(x) \notin B(y_i, r)\}} 0 \psi_i(f(x)) \in B(0, r) \end{aligned}$$

because $0 \in B(0, r)$, $B(0, r)$ is convex, and 6.6. It is just a convex combination of things in $B(0, r)$. ■

Note that we could have had the y_i **in** $f(K)$ **in addition to being in** $\overline{f(K)}$. This would make it possible to eliminate the assumption that K is closed later on. All you really need is that K is convex.

We think of f_r as an approximation to f. In fact it is uniformly within r of f on K. The next lemma shows that this f_r has a fixed point. This is the main result and comes from the Brouwer fixed point theorem in $\mathbb{R}^n$. This will be an approximate fixed point.

Lemma 6.4.2 *For each $r > 0$, there exists $x_r \in$ convex hull of $\overline{f(K)} \subseteq K$ such that*

$$f_r(x_r) = x_r, \ \|f_r(x) - f(x)\| < r \text{ for all } x$$

Proof: If $f_r(x_r) = x_r$ and $x_r = \sum_{i=1}^n a_i y_i$ for $\sum_{i=1}^n a_i = 1$ and the y_i described in the above lemma, we need

$$f_r(x_r) \equiv \sum_{i=1}^n y_i \psi_i(f(x_r)) = \sum_{j=1}^n y_j \psi_j \left(f\left(\sum_{i=1}^n a_i y_i \right) \right) = \sum_{j=1}^n a_j y_j = x_r.$$

Also, if this is satisfied, then we have the desired approximate fixed point.

This will be satisfied if for each $j = 1, \cdots, n$,

$$a_j = \psi_j \left(f\left(\sum_{i=1}^n a_i y_i \right) \right); \tag{6.8}$$

so, let

$$\Sigma_{n-1} \equiv \left\{ \boldsymbol{a} \in \mathbb{R}^n : \sum_{i=1}^n a_i = 1, \ a_i \geq 0 \right\}$$

and let $h : \Sigma_{n-1} \to \Sigma_{n-1}$ be given by

$$h(\boldsymbol{a})_j \equiv \psi_j \left(f\left(\sum_{i=1}^n a_i y_i \right) \right).$$

Since h is a continuous function of $\boldsymbol{a}$, the Brouwer fixed point theorem applies and there exists a fixed point for h which is a solution to 6.8. ■

The following is the Schauder fixed point theorem.

Theorem 6.4.3 *Let K be a closed and convex subset of X, a normed linear space. Let $f : K \to K$ be continuous and suppose $\overline{f(K)}$ is compact. Then f has a fixed point.*

Proof: Recall that $f(x_r) - f_r(x_r) \in B(0, r)$ and $f_r(x_r) = x_r$ with $x_r \in$ convex hull of $\overline{f(K)} \subseteq K$.

There is a subsequence, still denoted with subscript r with $r \to 0$ such that $f(x_r) \to x \in \overline{f(K)}$. **Note that the fact that K is convex is what makes f defined at x_r. x_r is in the convex hull of $\overline{f(K)} \subseteq K$. This is where we use K convex.** Then since f_r is uniformly close to f, it follows that $f(x_r) = x_r \to x$ also. Therefore,

$$f(x) = \lim_{r \to 0} f(x_r) = \lim_{r \to 0} f_r(x_r) = \lim_{r \to 0} x_r = x. \ ■$$

We usually have in mind the mapping defined on a Banach space. However, the completeness was never used. Thus the result holds in a normed linear space.

There is a nice corollary of this major theorem which is called the Schaefer fixed point theorem or the Leray Schauder alterative principle [22].

Theorem 6.4.4 *Let $f : X \to X$ be a compact map. Then either*

1. *There is a fixed point for f for all $t \in [0, 1]$ or*

2. *For every $r > 0$, there exists a solution to $x = tf(x)$ for $t \in (0,1)$ such that $\|x\| > r$.*

Proof: Suppose there is $t_0 \in [0,1]$ such that $t_0 f$ has no fixed point. Then $t_0 \neq 0$. If $t_0 = 0$, then $t_0 f$ obviously has a fixed point. Thus $t_0 \in (0,1]$. Then let r_M be the radial retraction onto $\overline{B(0,M)}$.

$$r_M f(x) = M \frac{f(x)}{\|f(x)\|}$$

By Schauder's theorem there exists $x \in \overline{B(0,M)}$ such that $t_0 r_M f(x) = x$. Then if $\|f(x)\| \leq M$, r_M has no effect and so $t_0 f(x) = x$ which is assumed not to take place. Hence $\|f(x)\| > M$ and so $\|r_M f(x)\| = M$ so $\|x\| = t_0 M$. Also $t_0 r_M f(x) = t_0 M \frac{f(x)}{\|f(x)\|} = x$ and so $x = \hat{t} f(x), \hat{t} = t_0 \frac{M}{\|f(x)\|} < 1$. Since M is arbitrary, it follows that the solutions to $x = tf(x)$ for $t \in (0,1)$ are unbounded. It was just shown that there is a solution to $x = \hat{t} f(x), \hat{t} < 1$ such that $\|x\| = t_0 M$ where M is arbitrary. Thus the second of the two alternatives holds. ■

As an example of the usefulness of the Schauder fixed point theorem, consider the following application to the theory of ordinary differential equations. In the context of this theorem, $X = C([0,T];\mathbb{R}^n)$, a Banach space with norm given by

$$\|\boldsymbol{x}\| \equiv \max\{|\boldsymbol{x}(t)| : t \in [0,T]\}.$$

I assume the reader knows about the Riemann integral in what follows and the elementary fundamental theorem of calculus. More general versions of these things are presented later in the book.

Theorem 6.4.5 *Let $\boldsymbol{f} : [0,T] \times \mathbb{R}^n \to \mathbb{R}^n$ be continuous and suppose there exists $L > 0$ such that for all $\lambda \in (0,1)$, if*

$$\boldsymbol{x}' = \lambda \boldsymbol{f}(t, \boldsymbol{x}),\ \boldsymbol{x}(0) = \boldsymbol{x}_0 \tag{6.9}$$

for all $t \in [0,T]$, then $\|\boldsymbol{x}\| < L$. Then there exists a solution to

$$\boldsymbol{x}' = \boldsymbol{f}(t, \boldsymbol{x}),\ \boldsymbol{x}(0) = \boldsymbol{x}_0 \tag{6.10}$$

for $t \in [0,T]$.

Proof: Let $F : X \to X$ where X described above.

$$F\boldsymbol{y}(t) \equiv \int_0^t \boldsymbol{f}(s, \boldsymbol{y}(s) + \boldsymbol{x}_0)\, ds$$

Let B be a bounded set in X. Then $|\boldsymbol{f}(s, \boldsymbol{y}(s) + \boldsymbol{x}_0)|$ is bounded for $s \in [0,T]$ if $\boldsymbol{y} \in B$. Say $|\boldsymbol{f}(s, \boldsymbol{y}(s) + \boldsymbol{x}_0)| \leq C_B$. Hence $F(B)$ is bounded in X. Also, for $\boldsymbol{y} \in B, s < t$,

$$|F\boldsymbol{y}(t) - F\boldsymbol{y}(s)| \leq \left| \int_s^t \boldsymbol{f}(s, \boldsymbol{y}(s) + \boldsymbol{x}_0)\, ds \right| \leq C_B |t - s|$$

and so $F(B)$ is pre-compact by the Ascoli Arzela theorem. By the Schaefer fixed point theorem, there are two alternatives. Either there are unbounded solutions $\boldsymbol{y}$ to

$$\lambda F(\boldsymbol{y}) = \boldsymbol{y}$$

for various $\lambda \in (0,1)$ or for all $\lambda \in [0,1]$, there is a fixed point for λF. In the first case, there would be unbounded $\boldsymbol{y}_\lambda$ solving

$$\boldsymbol{y}_\lambda(t) = \lambda \int_0^t \boldsymbol{f}(s, \boldsymbol{y}_\lambda(s) + \boldsymbol{x}_0)\, ds$$

Then let $\boldsymbol{x}_\lambda(s) \equiv \boldsymbol{y}_\lambda(s) + \boldsymbol{x}_0$ and you get $\|\boldsymbol{x}_\lambda\|$ also unbounded for various $\lambda \in (0,1)$. The above implies

$$\boldsymbol{x}_\lambda(t) - \boldsymbol{x}_0 = \lambda \int_0^t \boldsymbol{f}(s, \boldsymbol{x}_\lambda(s))\, ds$$

so $\boldsymbol{x}_\lambda' = \lambda \boldsymbol{f}(t, \boldsymbol{x}_\lambda), \boldsymbol{x}_\lambda(0) = \boldsymbol{x}_0$ and these would be unbounded for $\lambda \in (0,1)$ contrary to the assumption that there exists an estimate for these valid for all $\lambda \in (0,1)$. Hence the first alternative must hold and hence there is $\boldsymbol{y} \in X$ such that $F\boldsymbol{y} = \boldsymbol{y}$. Then letting $\boldsymbol{x}(s) \equiv \boldsymbol{y}(s) + \boldsymbol{x}_0$, it follows that

$$\boldsymbol{x}(t) - \boldsymbol{x}_0 = \int_0^t \boldsymbol{f}(s, \boldsymbol{x}(s))\, ds$$

and so $\boldsymbol{x}$ is a solution to the differential equation on $[0,T]$. ■

Note that existence of a solution to the differential equation is not assumed, only estimates of possible solutions. These estimates are called *a-priori* estimates. Also note this is a global existence theorem, not a local one for a solution defined on only a small interval.

6.5 The Kakutani Fixed Point Theorem

Definition 6.5.1 *If $A : X \to \mathscr{P}(Y)$ is a set-valued map, define the graph of A by*

$$G(A) \equiv \{(x,y) : y \in Ax\}.$$

Consider a map A which maps $\mathbb{C}^p$ to $\mathscr{P}(\mathbb{C}^p)$ which satisfies

$$A\boldsymbol{x} \text{ is compact and convex.} \tag{6.11}$$

and also the condition that if O is open and $O \supseteq A\boldsymbol{x}$, then there exists $\delta > 0$ such that if

$$\boldsymbol{y} \in B(\boldsymbol{x}, \delta), \text{ then } A\boldsymbol{y} \subseteq O. \tag{6.12}$$

*This last condition is sometimes referred to as **upper semicontinuity**. In words, A is upper semicontinuous and has values which are compact and convex. This is equivalent to saying that if $A\boldsymbol{x} \in O$ and $\boldsymbol{x}_n \to \boldsymbol{x}$, then for large enough n, it follows that $A\boldsymbol{x}_n \subseteq O$.*

With this definition, here is a lemma which has to do with the situation when the graph is closed.

Lemma 6.5.2 *Let A satisfy 6.12. Then AK is a subset of a compact set whenever K is compact. Also the graph of A is closed if $A\boldsymbol{x}$ is closed.*

Proof: Let $\boldsymbol{x} \in K$. Then $A\boldsymbol{x}$ is compact and contained in some open set whose closure is compact, $U_{\boldsymbol{x}}$. By assumption 6.12 there exists an open set $V_{\boldsymbol{x}}$ containing $\boldsymbol{x}$ such that if $\boldsymbol{y} \in V_{\boldsymbol{x}}$, then $A\boldsymbol{y} \subseteq U_{\boldsymbol{x}}$. Let $V_{\boldsymbol{x}_1}, \cdots, V_{\boldsymbol{x}_m}$ cover K. Then $AK \subseteq \cup_{k=1}^m \overline{U}_{\boldsymbol{x}_k}$, a compact set.

To see the graph of A is closed when $A\boldsymbol{x}$ is closed, let $\boldsymbol{x}_k \to \boldsymbol{x}, \boldsymbol{y}_k \to \boldsymbol{y}$ where $\boldsymbol{y}_k \in A\boldsymbol{x}_k$. Then letting $O = A\boldsymbol{x} + B(\boldsymbol{0}, r)$ it follows from 6.12 that $\boldsymbol{y}_k \in A\boldsymbol{x}_k \subseteq O$ for all k large enough. Therefore, $\boldsymbol{y} \in A\boldsymbol{x} + B(\boldsymbol{0}, 2r)$ and since $r > 0$ is arbitrary and $A\boldsymbol{x}$ is closed it follows $\boldsymbol{y} \in A\boldsymbol{x}$. ■

Also, there is a general consideration relative to upper semicontinuous functions.

Lemma 6.5.3 *If $\boldsymbol{f}$ is upper semicontinuous on some set K and $\boldsymbol{g}$ is continuous and defined on $\boldsymbol{f}(K)$, then $\boldsymbol{g}\circ\boldsymbol{f}$ is also upper semicontinuous.*

Proof: Let $\boldsymbol{x}_n \to \boldsymbol{x}$ in K. Let $U \supseteq \boldsymbol{g}\circ\boldsymbol{f}(\boldsymbol{x})$. Is $\boldsymbol{g}\circ\boldsymbol{f}(\boldsymbol{x}_n) \in U$ for all n large enough? We have $\boldsymbol{f}(\boldsymbol{x}) \in \boldsymbol{g}^{-1}(U)$, an open set. Therefore, if n is large enough, $\boldsymbol{f}(\boldsymbol{x}_n) \in \boldsymbol{g}^{-1}(U)$. It follows that for large enough n, $\boldsymbol{g}\circ\boldsymbol{f}(\boldsymbol{x}_n) \in U$ and so $\boldsymbol{g}\circ\boldsymbol{f}$ is upper semicontinuous on K. ■

The next theorem is an application of the Brouwer fixed point theorem. First define an p simplex, denoted by $[\boldsymbol{x}_0,\cdots,\boldsymbol{x}_p]$, to be the convex hull of the $p+1$ points, $\{\boldsymbol{x}_0,\cdots,\boldsymbol{x}_p\}$ where $\{\boldsymbol{x}_i-\boldsymbol{x}_0\}_{i=1}^p$ are independent. Thus

$$[\boldsymbol{x}_0,\cdots,\boldsymbol{x}_p] \equiv \left\{\sum_{i=1}^p t_i\boldsymbol{x}_i : \sum_{i=1}^p t_i = 1,\ t_i \geq 0\right\}.$$

If $p \leq 2$, the simplex is a triangle, line segment, or point. If $p \leq 3$, it is a tetrahedron, triangle, line segment or point. A collection of simplices is a tiling of $\mathbb{R}^p$ if $\mathbb{R}^p$ is contained in their union and if S_1, S_2 are two simplices in the tiling, with

$$S_j = \left[\boldsymbol{x}_0^j,\cdots,\boldsymbol{x}_p^j\right],$$

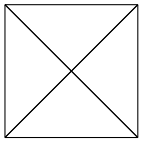
then

$$S_1 \cap S_2 = \left[\boldsymbol{x}_{k_0},\cdots,\boldsymbol{x}_{k_r}\right]$$

where

$$\left\{\boldsymbol{x}_{k_0},\cdots,\boldsymbol{x}_{k_r}\right\} \subseteq \left\{\boldsymbol{x}_0^1,\cdots,\boldsymbol{x}_p^1\right\} \cap \left\{\boldsymbol{x}_0^2,\cdots,\boldsymbol{x}_p^2\right\}$$

or else the two simplices do not intersect. The collection of simplices is said to be locally finite if, for every point, there exists a ball containing that point which also intersects only finitely many of the simplices in the collection. It is left to the reader to verify that for each $\varepsilon > 0$, there exists a locally finite tiling of $\mathbb{R}^p$ which is composed of simplices which have diameters less than ε. The local finiteness ensures that for each ε the vertices have no limit point. To see how to do this, consider the case of $\mathbb{R}^2$. Tile the plane with identical small squares and then form the triangles indicated in the following picture. It is clear something similar can be done in any dimension. Making the squares identical ensures that the little triangles are locally finite.



In general, you could consider $[0,1]^p$. The point at the center is $(1/2,\cdots,1/2)$. Then there are $2p$ faces. Form the $2p$ pyramids having this point along with the 2^{p-1} vertices of the face. Then use induction on each of these faces to form smaller dimensional simplices tiling that face. Corresponding to each of these $2p$ pyramids, it is the union of the simplices whose vertices consist of the center point along with those of these new simplicies tiling the chosen face. In general, you can write any p dimensional cube as the translate of a scaled $[0,1]^p$. Thus one can express each of identical cubes as a tiling of $m(p)$ simplices of the

appropriate size and thereby obtain a tiling of $\mathbb{R}^p$ with simplices. A ball will intersect only finitely many of the cubes and hence finitely many of the simplices. To get their diameters small as desired, just use $[0,r]^p$ instead of $[0,1]^p$.

Thus one can give a function any value desired on these vertices and extend appropriately to the rest of the simplex and obtain a continuous function.

The Kakutani fixed point theorem is a generalization of the Brouwer fixed point theorem from continuous single valued maps to upper semicontinuous maps which have closed convex values.

Theorem 6.5.4 *Let K be a compact convex subset of $\mathbb{R}^p$ and let $A : K \to \mathscr{P}(K)$ such that $A\boldsymbol{x}$ is a closed convex subset of K and A is upper semicontinuous. Then there exists $\boldsymbol{x}$ such that $\boldsymbol{x} \in A\boldsymbol{x}$. This is the "fixed point".*

Proof: Let there be a locally finite tiling of $\mathbb{R}^p$ consisting of simplices having diameter no more than ε. Let $P\boldsymbol{x}$ be the point in K which is closest to $\boldsymbol{x}$. For each vertex $\boldsymbol{x}_k$, pick $A_\varepsilon \boldsymbol{x}_k \in AP\boldsymbol{x}_k$ and define A_ε on all of $\mathbb{R}^p$ by the following rule. If

$$\boldsymbol{x} \in [\boldsymbol{x}_0, \cdots, \boldsymbol{x}_p],$$

so $\boldsymbol{x} = \sum_{i=0}^p t_i \boldsymbol{x}_i, t_i \in [0,1], \sum_i t_i = 1$,then

$$A_\varepsilon\, \boldsymbol{x} \equiv \sum_{k=0}^p t_k A_\varepsilon \boldsymbol{x}_k.$$

Now by construction $A_\varepsilon \boldsymbol{x}_k \in AP\boldsymbol{x}_k \in K$ and so A_ε is a continuous map defined on $\mathbb{R}^p$ with values in K thanks to the local finiteness of the collection of simplices. By the Brouwer fixed point theorem A_ε has a fixed point $\boldsymbol{x}_\varepsilon$ in K, $A_\varepsilon \boldsymbol{x}_\varepsilon = \boldsymbol{x}_\varepsilon$.

$$\boldsymbol{x}_\varepsilon = \sum_{k=0}^p t_k^\varepsilon A_\varepsilon \boldsymbol{x}_k^\varepsilon,\ A_\varepsilon \boldsymbol{x}_k^\varepsilon \in AP\boldsymbol{x}_k^\varepsilon \subseteq K$$

where a simplex containing $\boldsymbol{x}_\varepsilon$ is

$$[\boldsymbol{x}_0^\varepsilon, \cdots, \boldsymbol{x}_p^\varepsilon],\ \boldsymbol{x}_\varepsilon = \sum_{k=0}^p t_k^\varepsilon \boldsymbol{x}_k^\varepsilon$$

Also, $\boldsymbol{x}_\varepsilon \in K$ and is closer than ε to each $\boldsymbol{x}_k^\varepsilon$ so each $\boldsymbol{x}_k^\varepsilon$ is within ε of K. It follows that for each k, $\left|P\boldsymbol{x}_k^\varepsilon - \boldsymbol{x}_k^\varepsilon\right| < \varepsilon$ and so

$$\lim_{\varepsilon \to 0} |P\boldsymbol{x}_k^\varepsilon - \boldsymbol{x}_k^\varepsilon| = 0$$

By compactness of K, there exists a subsequence, still denoted with the subscript of ε such that for each k, the following convergences hold as $\varepsilon \to 0$

$$t_k^\varepsilon \to t_k,\ A_\varepsilon \boldsymbol{x}_k^\varepsilon \to \boldsymbol{y}_k,\ P\boldsymbol{x}_k^\varepsilon \to \boldsymbol{z}_k,\ \boldsymbol{x}_k^\varepsilon \to \boldsymbol{z}_k$$

Any pair of the $\boldsymbol{x}_k^\varepsilon$ are within ε of each other. Hence, any pair of the $P\boldsymbol{x}_k^\varepsilon$ are within ε of each other because P reduces distances. Therefore, in fact, $\boldsymbol{z}_k$ does not depend on k.

$$\lim_{\varepsilon \to 0} P\boldsymbol{x}_k^\varepsilon = \lim_{\varepsilon \to 0} \boldsymbol{x}_k^\varepsilon = \boldsymbol{z},\ \lim_{\varepsilon \to 0} \boldsymbol{x}_\varepsilon = \lim_{\varepsilon \to 0} \sum_{k=0}^p t_k^\varepsilon \boldsymbol{x}_k^\varepsilon = \sum_{k=0}^p t_k \boldsymbol{z} = \boldsymbol{z}$$

By upper semicontinuity of A, for all ε small enough,

$$AP\boldsymbol{x}_k^\varepsilon \subseteq A\boldsymbol{z} + B(\mathbf{0}, r)$$

In particular, since $A_\varepsilon \boldsymbol{x}_k^\varepsilon \in AP\boldsymbol{x}_k^\varepsilon$,

$$A_\varepsilon \boldsymbol{x}_k^\varepsilon \in A\boldsymbol{z} + B(\mathbf{0}, r) \text{ for } \varepsilon \text{ small enough}$$

Since r is arbitrary and $A\boldsymbol{z}$ is closed, it follows $\boldsymbol{y}_k \in A\boldsymbol{z}$. It follows that since K is closed,

$$\boldsymbol{x}_\varepsilon \to \boldsymbol{z} = \sum_{k=0}^{p} t_k \boldsymbol{y}_k,\ t_k \geq 0,\ \sum_{k=0}^{p} t_k = 1$$

Now by convexity of $A\boldsymbol{z}$ and the fact just shown that $\boldsymbol{y}_k \in A\boldsymbol{z}$,

$$\boldsymbol{z} = \sum_{k=0}^{p} t_k \boldsymbol{y}_k \in A\boldsymbol{z}$$

and so $\boldsymbol{z} \in A\boldsymbol{z}$. This is the fixed point. ■

One can replace $\mathbb{R}^p$ with $\mathbb{C}^p$ in the above theorem because it is essentially $\mathbb{R}^{2p}$. Also the theorem holds with no change for any finite dimensional normed linear space since these are homeomorpic to $\mathbb{R}^p$ or $\mathbb{C}^p$.

6.6 Ekeland's Variational Principle

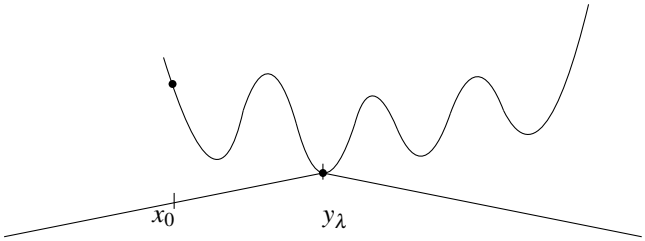
Recall the notation $X' = \mathscr{L}(X, \mathbb{R})$, the continuous linear functions mapping X to $\mathbb{R}$. This section deals with real Banach spaces. If you had complex ones, X' would denote $\mathscr{L}(X, \mathbb{C})$.

Definition 6.6.1 *A function $\phi : X \to (-\infty, \infty]$ is called proper if it is not constantly equal to ∞. Here X is assumed to be a complete metric space. The function ϕ is lower semicontinuous if*

$$x_n \to x \text{ implies } \phi(x) \leq \liminf_{n\to\infty} \phi(x_n)$$

It is bounded below if there is some constant C such that $C \leq \phi(x)$ for all x.

The variational principle of Ekeland is the following theorem [22]. You start with an approximate minimizer x_0. It says there is y_λ fairly close to x_0 such that if you subtract a "cone" from the value of ϕ at y_λ, then the resulting function is less than $\phi(x)$ for all $x \neq y_\lambda$.



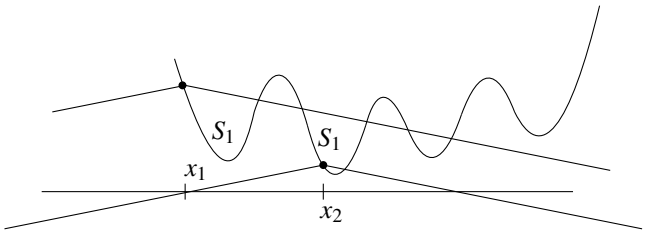
Theorem 6.6.2 *Let X be a complete metric space and let $\phi : X \to (-\infty, \infty]$ be proper, lower semicontinuous and bounded below. Let x_0 be such that*

$$\phi(x_0) \leq \inf_{x\in X} \phi(x) + \varepsilon$$

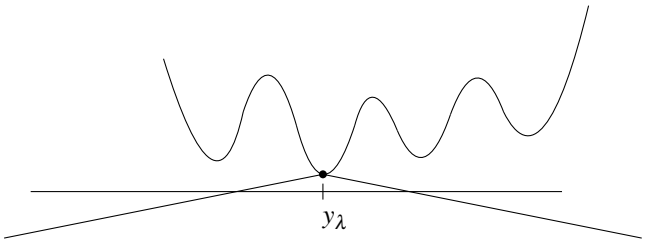
Then for every $\lambda > 0$ there exists a y_λ such that

1. $\phi(y_\lambda) \leq \phi(x_0)$
2. $d(y_\lambda, x_0) \leq \lambda$
3. $\phi(y_\lambda) - \frac{\varepsilon}{\lambda} d(x, y_\lambda) < \phi(x)$ *for all* $x \neq y_\lambda$

To motivate the proof, see the following picture which illustrates the first two steps. The $S_i$will be sets in X but are denoted symbolically by labeling them in $X \times (-\infty, \infty]$.



Then the end result of this iteration would be a picture like the following.



Thus you would have $\phi(y_\lambda) - \frac{\varepsilon}{\lambda} d(y_\lambda, x) \leq \phi(x)$ for all x which is seen to be what is wanted.

Proof: Let $x_1 = x_0$ and define $S_1 \equiv \left\{ z \in X : \phi(z) \leq \phi(x_1) - \frac{\varepsilon}{\lambda} d(z, x_1) \right\}$. Then S_1 contains x_1 so it is nonempty. It is also clear that S_1 is a closed set. This follows from the lower semicontinuity of ϕ. Suppose

$$S_k \equiv \left\{ z \in X : \phi(z) \leq \phi(x_k) - \frac{\varepsilon}{\lambda} d(z, x_k) \right\}$$

where $x_k \in S_{k-1}$. Pick $x_{k+1} \in S_k$ and define S_{k+1} similarly. Will this yield a nested sequence of nonempty closed sets? Yes, it appears that it would because if $z \in S_k$ then

$$\begin{aligned} \phi(z) &\leq \overset{\in S_{k-1}}{\phi(x_k)} - \frac{\varepsilon}{\lambda} d(z, x_k) \leq \left(\phi(x_{k-1}) - \frac{\varepsilon}{\lambda} d(x_{k-1}, x_k) \right) - \frac{\varepsilon}{\lambda} d(z, x_k) \\ &\leq \phi(x_{k-1}) - \frac{\varepsilon}{\lambda} d(z, x_{k-1}) \end{aligned}$$

showing that z has what it takes to be in S_{k-1}. Thus we would obtain a sequence of nested, nonempty, closed sets according to this scheme.

Now here is how to choose the $x_k \in S_{k-1}$. Let $\phi(x_k) < \inf_{x \in S_{k-1}} \phi(x) + \frac{1}{2^k}$. Then for $z \in S_{n+1} \subseteq S_n$, $\phi(z) \leq \phi(x_{n+1}) - \frac{\varepsilon}{\lambda} d(z, x_{n+1})$ and so

$$\begin{aligned} \frac{\varepsilon}{\lambda} d(z, x_{n+1}) &\leq \phi(x_{n+1}) - \phi(z) \leq \inf_{x \in S_n} \phi(x) + \frac{1}{2^{n+1}} - \phi(z) \\ &\leq \phi(z) + \frac{1}{2^{n+1}} - \phi(z) = \frac{1}{2^{n+1}} \end{aligned}$$

Thus every $z \in S_{n+1}$ is within $\frac{1}{2^{n+1}}$ of the single point x_{n+1} and so the diameter of S_n converges to 0 as $n \to \infty$. By completeness of X, there exists a unique $y_\lambda \in \cap_n S_n$. Then it follows in particular that for $x_0 = x_1$ as above, $\phi(y_\lambda) \leq \phi(x_0) - \frac{\varepsilon}{\lambda} d(y_\lambda, x_0) \leq \phi(x_0)$ which verifies the first of the above conclusions.

As to the second, $\phi(x_0) \leq \inf_{x \in X} \phi(x) + \varepsilon$ and so, for any x,

$$\phi(y_\lambda) \leq \phi(x_0) - \frac{\varepsilon}{\lambda} d(y_\lambda, x_0) \leq \phi(x) + \varepsilon - \frac{\varepsilon}{\lambda} d(y_\lambda, x_0),$$

this being true for $x = y_\lambda$. Hence $\frac{\varepsilon}{\lambda} d(y_\lambda, x_0) \leq \varepsilon$ and so $d(y_\lambda, x_0) \leq \lambda$.

Finally consider the third condition. If it does not hold, then there exists $z \neq y_\lambda$ such that $\phi(y_\lambda) \geq \phi(z) + \frac{\varepsilon}{\lambda} d(z, y_\lambda)$ so that $\phi(z) \leq \phi(y_\lambda) - \frac{\varepsilon}{\lambda} d(z, y_\lambda)$. But then, by the definition of y_λ as being in all the S_n, $\phi(y_\lambda) \leq \phi(x_n) - \frac{\varepsilon}{\lambda} d(x_n, y_\lambda)$ and so

$$\begin{aligned} \phi(z) &\leq \phi(x_n) - \frac{\varepsilon}{\lambda} \left(d(x_n, y_\lambda) + d(z, y_\lambda) \right) \\ &\leq \phi(x_n) - \frac{\varepsilon}{\lambda} d(x_n, z) \end{aligned}$$

Since n is arbitrary, this shows that $z \in \cap_n S_n$ but there is only one element of this intersection and it is y_λ so z must equal y_λ, a contradiction. ■

Note how if you make λ very small, you could pick ε very small such that the cone looks pretty flat. Of course, you can always consider an equivalent metric $\hat{d}(x,y) \equiv \frac{\varepsilon}{\lambda} d(x,y)$ in all of these considerations.

6.6.1 Cariste Fixed Point Theorem

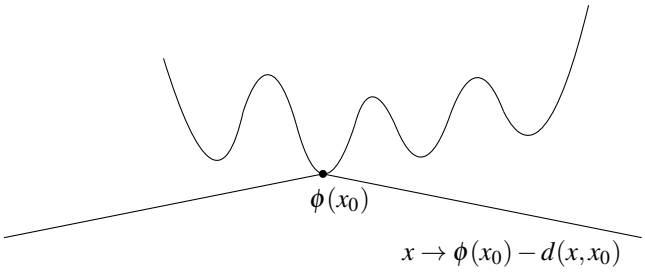
As mentioned in [22], the above result can be used to prove the Cariste fixed point theorem.

Theorem 6.6.3 *Let ϕ be lower semicontinuous, proper, and bounded below on a complete metric space X and let $F : X \to \mathscr{P}(X)$ be set valued such that $F(x) \neq \emptyset$ for all x. Also suppose that for each $x \in X$, there exists $y \in F(x)$ such that $\phi(y) \leq \phi(x) - d(x,y)$. Then there exists x_0 such that $x_0 \in F(x_0)$.*

Proof: In the above Ekeland variational principle, let $\varepsilon = 1 = \lambda$. Then there exists x_0 such that for all $y \neq x_0$

$$\phi(x_0) - d(y, x_0) < \phi(y), \text{ so } \phi(x_0) < \phi(y) + d(y, x_0) \tag{6.13}$$

for all $y \neq x_0$.



Suppose $x_0 \notin F(x_0)$. From the assumption, there is $y \in F(x_0)$ (so $y \neq x_0$) such that $\phi(y) \leq \phi(x_0) - d(x_0, y)$. Since $y \neq x_0$, it follows

$$\phi(y) + d(x_0, y) \leq \phi(x_0) < \phi(y) + d(y, x_0)$$

a contradiction. Hence $x_0 \in F(x_0)$ after all. ■

It is a funny theorem. It is easy to prove, but you look at it and wonder what it says. In fact, it implies the Banach fixed point theorem. If F is single valued, you would need to have a function ϕ such that for each x,

$$\phi(F(x)) \leq \phi(x) - d(x, y)$$

and if you have such a ϕ then you can assert there is a fixed point for F. Suppose F is single valued and $d(Fx, Fy) \leq r d(x, y), 0 < r < 1$. Of course F has a fixed point using easier techniques. However, this also follows from this result. Let $\phi(x) = \frac{1}{1-r} d(x, F(x))$. Then is it true that for each x, there exists $y \in F(x)$ such that the inequality holds for all x? Is

$$\frac{1}{1-r} d(F(x), F(F(x))) \leq \frac{1}{1-r} d(x, F(x)) - d(x, F(x))$$

Yes, this is certainly so because the right side reduces to $\frac{r}{1-r} d(x, F(x))$. Thus this fixed point theorem implies the usual Banach fixed point theorem.

The Ekeland variational principle says that when ϕ is lower semicontinuous proper and bounded below, there exists y such that

$$\phi(y) - d(x, y) < \phi(x) \text{ for all } x \neq y$$

In fact this can be proved from the Cariste fixed point theorem. Suppose the variational principle does not hold. This would mean that for all y there exists $x \neq y$ such that $\phi(y) - d(x, y) \geq \phi(x)$. Thus, for all x there exists $y \neq x$ such that $\phi(x) - d(x, y) \geq \phi(y)$. The inequality is preserved if $x = y$. Then let

$$F(x) \equiv \{y \neq x : \phi(x) - d(x, y) \geq \phi(y)\} \neq \emptyset$$

by assumption. This is the hypothesis for the Cariste fixed point theorem. Hence there exists x_0 such that $x_0 \in F(x_0) = \{y \neq x_0 : \phi(x_0) - d(x_0, y) \geq \phi(y)\}$ but this cannot happen because you can't have $x_0 \neq x_0$. Thus the Ekeland variational principle must hold after all.

6.6.2 A Density Result

There are several applications of the Ekeland variational principle. For more of them, see [22]. One of these is to show that there is a point where ϕ' is small assuming ϕ is bounded below, lower semicontinuous, and Gateaux differentiable, meaning that there exists $\phi'(x) \in X'$ such that if $v \in X$, then

$$\phi'(x)(v) \equiv \lim_{h \to 0} \frac{\phi(x + hv) - \phi(x)}{h}, \ \phi'(x) \in X'$$

Here X is a real Banach space.

Theorem 6.6.4 *Let X be a Banach space and $\phi : X \to \mathbb{R}$ be Gateaux differentiable, bounded from below, and lower semicontinuous. Then for every $\varepsilon > 0$ there exists $x \in X$ such that*

$$\phi(x_\varepsilon) \leq \inf_{x \in X} \phi(x) + \varepsilon \text{ and } \left\| \phi'(x_\varepsilon) \right\|_{X'} \leq \varepsilon$$

Proof: From the Ekeland variational principle with $\lambda = 1$, there exists x_ε such that $\phi(x_\varepsilon) \leq \phi(x_0) \leq \inf_{x\in X} \phi(x) + \varepsilon$ and for all x, $\phi(x_\varepsilon) < \phi(x) + \varepsilon \|x - x_\varepsilon\|$. Then letting $x = x_\varepsilon + hv$ where $\|v\| = 1, \phi(x_\varepsilon + hv) - \phi(x_\varepsilon) > -\varepsilon |h|$. Let $h < 0$. Then divide by it to obtain $\frac{\phi(x_\varepsilon+hv)-\phi(x_\varepsilon)}{h} < \varepsilon$. Passing to a limit as $h \to 0$ yields $\phi'(x)(v) \leq \varepsilon$. Now v was arbitrary with norm 1 and so $\sup_{\|v\|=1} |\phi'(x_\varepsilon)(v)| = \|\phi'(x_\varepsilon)\| \leq \varepsilon$ ■

There is another very interesting application of the Ekeland variational principle [22].

Theorem 6.6.5 *Let X be a real Banach space and $\phi : X \to \mathbb{R}$ be Gateaux differentiable, bounded from below, and lower semicontinuous. Also suppose there exists $a, c > 0$ such that*

$$a\|x\| - c \leq \phi(x) \text{ for all } x \in X$$

Then $\{\phi'(x) : x \in X\}$ is dense in the ball of X' centered at 0 with radius a. Here $\phi'(x) \in X'$ and is determined by

$$\phi'(x)(v) \equiv \lim_{h\to 0} \frac{\phi(x+hv) - \phi(x)}{h}$$

Proof: Let $x^* \in X', \|x^*\| \leq a$. Let $\psi(x) = \phi(x) - x^*(x)$. This is lower semicontinuous. It is also bounded from below because

$$\psi(x) \geq \phi(x) - a\|x\| \geq (a\|x\| - c) - a\|x\| = -c$$

It is also clearly Gateaux differentiable and lower semicontinuous because the piece added in is actually continuous. It is clear that the Gateaux derivative is just $\phi'(x) - x^*$. By Theorem 6.6.4, there exists x_ε such that $\|\phi'(x_\varepsilon) - x^*\| \leq \varepsilon$ ■

Thus this theorem says that if $\phi(x) \geq a\|x\| - c$ where ϕ has the nice properties of the theorem, it follows that $\phi'(x)$ is dense in $B(0,a)$ in the dual space X'. It follows that if for every a, there exists c such that $\phi(x) \geq a\|x\| - c$ for all $x \in X$ then $\{\phi'(x) : x \in X\}$ is dense in X'. This proves the following lemma.

Lemma 6.6.6 *Let X be a real Banach space and $\phi : X \to \mathbb{R}$ be Gateaux differentiable, bounded from below, and lower semicontinuous. Suppose for all $a > 0$ there exists a $c > 0$ such that $\phi(x) \geq a\|x\| - c$ for all x. Then $\{\phi'(x) : x \in X\}$ is dense in X'.*

If the above holds, then $\frac{\phi(x)}{\|x\|} \geq a - \frac{c}{\|x\|}$ and so, since a is arbitrary, it must be the case that

$$\lim_{\|x\|\to\infty} \frac{\phi(x)}{\|x\|} = \infty. \tag{6.14}$$

In fact, this is sufficient to conclude that for each $a > 0$ there is $c > 0$ such that $\phi(x) \geq a\|x\| - c$. If not, there would exist $a > 0$ such that $\phi(x_n) < a\|x_n\| - n$. Let $-L$ be a lower bound for $\phi(x)$. Then $-L + n \leq a\|x_n\|$ and so $\|x_n\| \to \infty$. Now it follows that

$$a \geq \frac{\phi(x_n)}{\|x_n\|} + \frac{n}{\|x_n\|} \geq \frac{\phi(x_n)}{\|x_n\|} \tag{6.15}$$

which is a contradiction to 6.14. This proves the following interesting density theorem.

Theorem 6.6.7 *Let X be a real Banach space and $\phi : X \to \mathbb{R}$ be Gateaux differentiable, bounded from below, and lower semicontinuous. Also suppose the coercivity condition*

$$\lim_{\|x\|\to\infty} \frac{\phi(x)}{\|x\|} = \infty$$

Then $\{\phi'(x) : x \in X\}$ *is dense in* X'. *Here* $\phi'(x) \in X'$ *and is determined by*

$$\phi'(x)(v) \equiv \lim_{h\to 0} \frac{\phi(x+hv) - \phi(x)}{h}$$

6.7 Exercises

1. It was shown that in a finite dimensional normed linear space that the compact sets are exactly those which are closed and bounded. Explain why every finite dimensional normed linear space is complete.

2. In any normed linear space, show that $\operatorname{span}(x_1, \cdots, x_n)$ is closed. That is, the span of any finite set of vectors is always a closed subspace. **Hint:** Suppose you let $V = \operatorname{span}(x_1, \cdots, x_n)$ and let $v^n \to v$ be convergent sequence of vectors in V. What does this say about the coordinate maps? Remember these are linear maps into $\mathbb{F}$ and so they are continuous.

3. It was shown that in a finite dimensional normed linear space that the compact sets are exactly those which are closed and bounded. What if you have an infinite dimensional normed linear space X? Show that the unit ball $D(0,r) \equiv \{x : \|x\| \leq 1\}$ is **NEVER** compact even though it is closed and bounded. **Hint:** Suppose you have $\{x_i\}_{i=1}^n$ where $\|x_i - x_j\| \geq \frac{1}{2}$. Let $y \notin \operatorname{span}(x_1, \cdots, x_n)$, a closed subspace. Such a y exists because X is not finite dimensional. Explain why $\operatorname{dist}(y, \operatorname{span}(x_1, \cdots, x_n)) > 0$. This depends on $\operatorname{span}(x_1, \cdots, x_n)$ being closed. Let $z \in \operatorname{span}(x_1, \cdots, x_n)$ such that $\|y - z\| \leq 2\operatorname{dist}(y, \operatorname{span}(x_1, \cdots, x_n))$. Let $x_{n+1} \equiv \frac{y-z}{\|y-z\|}$. Then consider the following:

$$\|x_{n+1} - x_k\| = \left\| \frac{y - (z + \|y-z\| x_k)}{\|y-z\|} \right\| \geq \frac{\|y - (z + \|y-z\| x_k)\|}{2\operatorname{dist}(y, \operatorname{span}(x_1, \cdots, x_n))}$$

   What of $(z + \|y-z\| x_k)$? Where is it? Isn't it in $\operatorname{span}(x_1, \cdots, x_n)$? Explain why this yields a sequence of points of X which are spaced at least 1/2 apart even though they are all in the closed unit ball.

4. Find an example of two 2×2 matrices A, B such that $\|AB\| < \|A\| \|B\|$. This refers to the operator norm taken with respect to the usual norm on $\mathbb{R}^2$. **Hint:** Maybe make it easy on yourself and consider diagonal matrices.

5. Now let $V = C([0,1])$ and let $T : V \to V$ be given by $Tf(x) \equiv \int_0^x f(t)\, dt$. Show that T is continuous and linear. Here the norm is

$$\|f\| \equiv \max\{|f(x)| : x \in [0,1]\}.$$

   Can you find $\|T\|$ where this is the operator norm defined by analogy to what was given in the chapter?

6. Show that in any metric space (X,d), if U is an open set and if $x \in U$, then there exists $r > 0$ such that the closure of $B(x,r)$, $\overline{B(x,r)} \subseteq U$. This says, in topological terms, that (X,d) is regular. Is it always the case in a metric space that $\overline{B(x,r)} = \{y : d(y,x) \leq r\} \equiv D(0,r)$? Prove or disprove. **Hint:** In fact, the answer to the last question is no.

7. Let (X,d) be a complete metric space. Let $\{U_n\}$ be a sequence of dense open sets. This means that $B(x,r)\cap U_n \neq \emptyset$ for every $x\in X$, and $r>0$. You know that $\cap_n U_n$ is not necessarily open. Show that it is nevertheless, dense. **Hint:** Let $D=\cap_n U_n$. You need to show that $B(x,r)\cap D\neq\emptyset$. There is a point $p_1\in U_1\cap B(x,r)$. Then there exists $r_1<1/2$ such that $B(p_1,r_1)\subseteq U_1\cap B(x,r)$. From the above problem, you can adjust r_1 such that $\overline{B(p_1,r_1)}\subseteq U_1\cap B(x,r)$. Next there exists $p_2\in B(p_1,r_1)\cap U_2$. Let $r_2<1/2^2$ be such that $\overline{B(p_2,r_2)}\subseteq B(p_1,r_1)\cap U_2\cap U_1$. Continue this way. You get a nested sequence of closed sets $\{B_k\}$ such that the diameter of B_k is no more than $1/2^{k-1}$, the k^{th} being contained in $B(p_{k-1},r_{k-1})\cap\cap_{i=1}^{k-1}U_i$. Explain why there is a unique point in the intersection of these closed sets which is in $B(x,r)\cap\cap_{k=1}^{\infty}U_k$. Then explain why this shows that D is dense.

8. The countable intersection of open sets is called a G_δ set. Show that the rational numbers $\mathbb{Q}$ is **NOT** a G_δ set in $\mathbb{R}$. In fact, show that no countable dense set can be a G_δ set. Show that $\mathbb{N}$ is a G_δ set. It is not dense.

9. You have a function $f:(X,d)\to(Y,\rho)$. Define

$$\omega_\delta f(x)\equiv\sup\{\rho(f(z),f(y)):z,y\in B(x,\delta)\}$$

Then explain why $\lim_{\delta\to0}\omega_\delta f(x)\equiv\omega f(x)$ exists. Explain why a function is continuous at x if and only if $\omega f(x)=0$. Next show that the set of all x where $\omega f(x)=0$ is a G_δ set. **Hint:** $\omega f(x)=0$ if and only if x is in something like this: $\cap_{n=1}^{\infty}\cup_{k=1}^{\infty}\left[\omega_{1/k}f(\omega)<\frac{1}{n}\right]$. Explain this. Then explain why $\cup_{k=1}^{\infty}\left[\omega_{1/k}f(\omega)<\frac{1}{n}\right]$ is an open set.

10. Prove or disprove.

    (a) If A is compact, then $\mathbb{R}^n\setminus A$ is connected. You might consider the case $n>1$ and the case $n=1$ separately.

    (b) If A is connected in $\mathbb{R}^n$, then $\mathbb{R}^n\setminus A$ is also connected.

    (c) If A is connected in $\mathbb{R}^n$, then either A is open or A is closed.

    (d) $\mathbb{R}^n\setminus B(\mathbf{0},1)$ is connected. Two cases to consider: $n=1$ and $n>1$.

11. If A is a connected set in $\mathbb{R}^n$, and A is not a single point, show that every point of A is a limit point of A.

12. Consider the Cantor set. This is obtained by starting with $[0,1]$ deleting $(1/3,2,3)$ and then taking the two closed intervals which result and deleting the middle open third of each of these and continuing this way. Let J_k denote the union of the 2^k closed intervals which result at the k^{th} step of the construction. The Cantor set is $J\equiv\cap_{k=1}^{\infty}J_k$. Explain why J is a nonempty compact subset of $\mathbb{R}$. Show that every point of J is a limit point of J. Also show there exists a mapping from J onto $[0,1]$ even though the sum of the lengths of the deleted open intervals is 1. Show that the Cantor set has empty interior. If $x\in J$, consider the connected component of x. Show that this connected component is just x.

13. You have a complete metric space (X,d) and a mapping $T:X\to X$ which satisfies $d(Tx,Ty)\leq rd(x,y)$, $0<r<1$. Show $x,Tx,T^2x,\cdots$. converges to a point $z\in X$ such that $Tz=z$. Next suppose you only know $\frac{d(Tx,x)}{1-r}<R$ and that on

$B(x,R), d(Tx,Ty) \leq rd(x,y)$ where $r<1$ as above. Show that then $z \in B(x,R)$ and that in fact each $T^k x \in B(x,R)$. Show also there is no more than one such fixed point z on $B(x,R)$.

14. In Theorem 5.6.1 it is assumed f has values in $\mathbb{F}$. Show there is no change if f has values in V, a normed vector space provided you redefine the definition of a polynomial to be something of the form $\sum_{|\alpha|\leq m} a_\alpha \boldsymbol{x}^\alpha$ where $a_\alpha \in V$.

15. How would you generalize the conclusion of Corollary 5.7.8 to include the situation where f has values in a finite dimensional normed vector space?

16. If f and g are real valued functions which are continuous on some set, D, show that $\min(f,g), \max(f,g)$ are also continuous. Generalize this to any finite collection of continuous functions. **Hint:** Note $\max(f,g) = \frac{|f-g|+f+g}{2}$. Now recall the triangle inequality which can be used to show $|\cdot|$ is a continuous function.

17. Find an example of a sequence of continuous functions defined on $\mathbb{R}^n$ such that each function is nonnegative and each function has a maximum value equal to 1 but the sequence of functions converges to 0 pointwise on $\mathbb{R}^n \setminus \{\mathbf{0}\}$, that is, the set of vectors in $\mathbb{R}^n$ excluding $\mathbf{0}$.

18. An open subset U of $\mathbb{R}^n$ is arcwise connected if and only if U is connected. Consider the usual Cartesian coordinates relative to axes $x_1, \cdots, x_n$. A square curve is one consisting of a succession of straight line segments each of which is parallel to some coordinate axis. Show an open subset U of $\mathbb{R}^n$ is connected if and only if every two points can be joined by a square curve.

19. Let $\boldsymbol{x} \to h(\boldsymbol{x})$ be a bounded continuous function. Show f is continuous for $f(\boldsymbol{x}) = \sum_{n=1}^{\infty} \frac{h(n\boldsymbol{x})}{n^2}$.

20. Let S be a any countable subset of $\mathbb{R}^n$. Show there exists a function, f defined on $\mathbb{R}^n$ which is discontinuous at every point of S but continuous everywhere else. **Hint:** This is real easy if you do the right thing. It involves the Weierstrass M test.

21. If f is any continuous function defined on K a sequentially compact subset of $\mathbb{R}^n$, show there exists a series of the form $\sum_{k=1}^{\infty} p_k$, where each p_k is a polynomial, which converges uniformly to f on $[a,b]$. **Hint:** You should use the Weierstrass approximation theorem to obtain a sequence of polynomials. Then arrange it so the limit of this sequence is an infinite sum.

22. Let K be a sequentially compact set in a normed vector space V and let $\boldsymbol{f}: V \to W$ be continuous where W is also a normed vector space. Show $\boldsymbol{f}(K)$ is also sequentially compact.

23. If $\boldsymbol{f}$ is uniformly continuous, does it follow that $|\boldsymbol{f}|$ is also uniformly continuous? If $|\boldsymbol{f}|$ is uniformly continuous does it follow that $\boldsymbol{f}$ is uniformly continuous? Answer the same questions with "uniformly continuous" replaced with "continuous". Explain why.

24. Suppose S, T are linear maps on some finite dimensional vector space, S^{-1} exists and let $\delta \in (0,1)$. Then whenever $\|S-T\|$ is small enough, it follows that

$$\frac{|T\boldsymbol{v}|}{|S\boldsymbol{v}|} \in (1-\delta, 1+\delta) \tag{6.16}$$

for all $\boldsymbol{v} \neq \boldsymbol{0}$. Similarly if T^{-1} exists and $\|S-T\|$ is small enough,

$$\frac{|T\boldsymbol{v}|}{|S\boldsymbol{v}|} \in (1-\delta, 1+\delta).$$

Hint: For the first part, consider the new norm $\|\boldsymbol{v}\| \equiv \left|S^{-1}\boldsymbol{v}\right|$. Use equivalence of norms and simple estimates to establish 6.16.

25. Let σ be an r simplex. Then $[\sigma, \boldsymbol{b}]$ will consist of all $(1-\lambda)\sigma + \lambda \boldsymbol{b}$ where $\lambda \in [0,1]$. If $\sigma = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_r]$, show that $[\sigma, \boldsymbol{b}] = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_r, \boldsymbol{b}]$. Now if $\sigma_1, \sigma_2 \subseteq \sigma$ where $[\sigma, \boldsymbol{b}]$ is an $r+1$ simplex and each σ_i is an r simplex, show that $[\sigma_1, \boldsymbol{b}] \cap [\sigma_2, \boldsymbol{b}] = [\sigma_2 \cap \sigma_1, \boldsymbol{b}]$.

26. Let $A : \mathbb{R}^n \to \mathbb{R}^n$ be continuous and let $\boldsymbol{f} \in \mathbb{R}^n$. Also let $(\cdot, \cdot)$ denote the standard inner product in $\mathbb{R}^n$. Letting K be a closed and bounded and convex set, show that there exists $\boldsymbol{x} \in K$ such that for all $\boldsymbol{y} \in K$, $(\boldsymbol{f} - A\boldsymbol{x}, \boldsymbol{y} - \boldsymbol{x}) \leq 0$. **Hint:** Show that this is the same as saying $P(\boldsymbol{f} - A\boldsymbol{x} + \boldsymbol{x}) = \boldsymbol{x}$ for some $\boldsymbol{x} \in K$ where here P is the projection map discussed above in Problem 10 on Page 152. Now use the Brouwer fixed point theorem. This little observation is called Browder's lemma. It is a fundamental result in nonlinear analysis.

27. ↑In the above problem, suppose that you have a coercivity result which is

$$\lim_{\|\boldsymbol{x}\| \to \infty} \frac{(A\boldsymbol{x}, \boldsymbol{x})}{\|\boldsymbol{x}\|} = \infty.$$

Show that if you have this, then you don't need to assume the convex closed set is bounded. In case $K = \mathbb{R}^n$, and this coercivity holds, show that A maps onto $\mathbb{R}^n$.

28. Let $f : X \to [-\infty, \infty]$ where X is a Banach space. This is said to be lower semicontinuous if whenever $x_n \to x$, it follows that $f(x) \leq \liminf_{n\to\infty} f(x_n)$. Show that this is the same as saying that the epigraph of f is closed. Here we can make $X \times [-\infty, \infty]$ into a metric space in a natural way by using the product topology where the distance on $[-\infty, \infty]$ will be $d(\sigma, \alpha) \equiv |\arctan(\sigma) - \arctan(\alpha)|$. Here $\text{epi}(f) \equiv \{(x, \alpha) : \alpha \geq f(x)\}$. The function is upper semicontinuous if $\limsup_{n\to\infty} f(x_n) \leq f(x)$. What is a condition for f to be upper semicontinuous? Do you need a Banach space to do this? Would it be sufficient to let X be a metric space?

29. Explain why the supremum of lower semicontinuous functions is lower semicontinuous and the infimum of upper semicontinuous functions is upper semicontinuous.

30. Let K be a nonempty closed and convex subset of $\mathbb{R}^n$. Recall K is convex means that if $\boldsymbol{x}, \boldsymbol{y} \in K$, then for all $t \in [0,1]$, $t\boldsymbol{x} + (1-t)\boldsymbol{y} \in K$. Show that if $\boldsymbol{x} \in \mathbb{R}^n$ there exists a unique $\boldsymbol{z} \in K$ such that $|\boldsymbol{x} - \boldsymbol{z}| = \min\{|\boldsymbol{x} - \boldsymbol{y}| : \boldsymbol{y} \in K\}$. This $\boldsymbol{z}$ will be denoted as $P\boldsymbol{x}$. **Hint:** First note you do not know K is compact. Establish the parallelogram

identity if you have not already done so, $|\boldsymbol{u}-\boldsymbol{v}|^2+|\boldsymbol{u}+\boldsymbol{v}|^2=2|\boldsymbol{u}|^2+2|\boldsymbol{v}|^2$. Then let $\{\boldsymbol{z}_k\}$ be a minimizing sequence,

$$\lim_{k\to\infty}|\boldsymbol{z}_k-\boldsymbol{x}|^2=\inf\{|\boldsymbol{x}-\boldsymbol{y}|:\boldsymbol{y}\in K\}\equiv\lambda.$$

Now using convexity, explain why

$$\left|\frac{\boldsymbol{z}_k-\boldsymbol{z}_m}{2}\right|^2+\left|\boldsymbol{x}-\frac{\boldsymbol{z}_k+\boldsymbol{z}_m}{2}\right|^2=2\left|\frac{\boldsymbol{x}-\boldsymbol{z}_k}{2}\right|^2+2\left|\frac{\boldsymbol{x}-\boldsymbol{z}_m}{2}\right|^2$$

and then use this to argue $\{\boldsymbol{z}_k\}$ is a Cauchy sequence. Then if $\boldsymbol{z}_i$ works for $i=1,2$, consider $(\boldsymbol{z}_1+\boldsymbol{z}_2)/2$ to get a contradiction.

31. In Problem 30 show that $P\boldsymbol{x}$ satisfies and is in fact characterized as the solution to the following variational inequality. $(\boldsymbol{x}-P\boldsymbol{x},\boldsymbol{y}-P\boldsymbol{x})\leq 0$ for all $\boldsymbol{y}\in K$. Then show that $|P\boldsymbol{x}_1-P\boldsymbol{x}_2|\leq|\boldsymbol{x}_1-\boldsymbol{x}_2|$. **Hint:** For the first part note that if $\boldsymbol{y}\in K$, the function $t\to|\boldsymbol{x}-(P\boldsymbol{x}+t(\boldsymbol{y}-P\boldsymbol{x}))|^2$ achieves its minimum on $[0,1]$ at $t=0$. For the second part,
$$(\boldsymbol{x}_1-P\boldsymbol{x}_1)\cdot(P\boldsymbol{x}_2-P\boldsymbol{x}_1)\leq 0,\ (\boldsymbol{x}_2-P\boldsymbol{x}_2)\cdot(P\boldsymbol{x}_1-P\boldsymbol{x}_2)\leq 0.$$
Explain why $(\boldsymbol{x}_2-P\boldsymbol{x}_2-(\boldsymbol{x}_1-P\boldsymbol{x}_1))\cdot(P\boldsymbol{x}_2-P\boldsymbol{x}_1)\geq 0$ and then use a some manipulations and the Cauchy Schwarz inequality to get the desired inequality. Thus P is called a retraction onto K.

32. Browder's lemma says: Let K be a convex closed and bounded set in $\mathbb{R}^n$ and let $A:K\to\mathbb{R}^n$ be continuous and $\boldsymbol{f}\in\mathbb{R}^n$. Then there exists $\boldsymbol{x}\in K$ such that for all $\boldsymbol{y}\in K$,
$$(\boldsymbol{f}-A\boldsymbol{x},\boldsymbol{y}-\boldsymbol{x})\leq 0$$
show this is true. **Hint:** Consider $\boldsymbol{x}\to P(\boldsymbol{f}-A\boldsymbol{x}+\boldsymbol{x})$ where P is the projection onto K. If there is a fixed point of this mapping, then $P(\boldsymbol{f}-A\boldsymbol{x}+\boldsymbol{x})=\boldsymbol{x}$. Now consider the variational inequality satisfied. This little lemma is the basis for a whole lot of nonlinear analysis involving nonlinear operators of various kinds.

33. Generalize the above problem as follows. Let K be a convex closed and bounded set in $\mathbb{R}^n$ and let $A:K\to\mathscr{P}(\mathbb{R}^n)$ be upper semi-continuous having closed bounded convex values and $\boldsymbol{f}\in\mathbb{R}^n$. Then there exists $\boldsymbol{x}\in K$ and $\boldsymbol{z}\in A\boldsymbol{x}$ such that for all $\boldsymbol{y}\in K$, $(\boldsymbol{f}-\boldsymbol{z},\boldsymbol{y}-\boldsymbol{x})\leq 0$ show this is true. Also show that if K is a closed convex and bounded set in E a finite dimensional normed linear space and $A:K\to\mathscr{P}(E')$ is upper semicontinuous having closed bounded convex values and $f\in E'$, then there exists $x\in K$ and $z\in Ax$ such that for all $y\in K$, $\langle f-z,y-x\rangle\leq 0$. **Hint:** Use the construction for the proof of the Kakutani fixed point theorem and the above Browder's lemma.

34. This problem establishes a remarkable result about existence for a system of inequalities based on the min max theorem, Theorem 5.11.5. Let E be a finite dimensional Banach space and let K be a convex and compact subset of E. A set valued map $A:D(A)\subseteq K\to E'$ is called monotone if whenever $v_i\in Au_i$, it follows that $\langle v_1-v_2,u_1-u_2\rangle\geq 0$. The graph, denoted as $\mathscr{G}(A)$ consists of all pairs $[u,v]$ such that $v\in Au$. This is a monotone subset of $E\times E'$. Let $z\in E'$ be fixed. Show that

for $[u_i, v_i] \in \mathscr{G}(A)$, for $i = 1, 2, \cdots, n$ there exists a solution $x \in K$ to the system of inequalities

$$\langle z + v_i, u_i - x \rangle \geq 0, i = 1, 2, \cdots, n$$

Hint: Let P_n be all $\vec{\lambda} = (\lambda_1, \cdots, \lambda_n)$ such that each $\lambda_k \geq 0$ and $\sum_{k=1}^n \lambda_k = 1$. Let $H : P_n \times P_n \to \mathbb{R}$ be given by

$$H\left(\vec{\mu}, \vec{\lambda}\right) \equiv \sum_{i=1}^n \mu_i \left\langle z + v_i, \sum_{j=1}^n \lambda_j u_j - u_i \right\rangle \tag{6.17}$$

Show that it is both convex and concave in both arguments. Then apply the min max theorem. Then argue that $H\left(\vec{\lambda}, \vec{\lambda}\right) \leq 0$ from monotonicity considerations. Letting $\left(\vec{\mu}_0, \vec{\lambda}_0\right)$ be the saddle point, you will have

$$\begin{aligned} H\left(\vec{\mu}, \vec{\lambda}_0\right) &\leq H\left(\vec{\mu}_0, \vec{\lambda}_0\right) \leq H\left(\vec{\mu}_0, \vec{\lambda}\right) \\ H\left(\vec{\mu}, \vec{\lambda}_0\right) &\leq H\left(\vec{\mu}_0, \vec{\lambda}_0\right) \leq H\left(\vec{\mu}_0, \vec{\mu}_0\right) \leq 0 \\ H\left(\vec{\mu}, \vec{\lambda}_0\right) &\leq 0 \end{aligned}$$

Now choose $\vec{\mu}$ judiciously while allowing $\vec{\lambda}_0$ to be used to define x which satisfies all the inequalities.

35. ↑It gets even better. Let $K_{u,v} \equiv \{x \in K : \langle z + v, u - x \rangle \geq 0\}$. Show that $K_{u,v}$ is compact and that the sets $K_{u,v}$ have the finite intersection property. Therefore, there exists $x \in \cap_{[u,v] \in \mathscr{G}(A)} K_{u,v}$. Explain why $\langle z + v, u - x \rangle \geq 0$ for all $[u, v] \in \mathscr{G}(A)$. What would the inequalities be if $-A$ were monotone?

36. Problem 33 gave a solution to the inequality $\langle f - z, y - x \rangle \leq 0, z \in Ax$ under the condition that A is upper semicontinuous. What are the differences between the result in the above problem and the result of Problem 33. You could replace A with $-A$ in the earlier problem. If you did, would you get the result of the above problem?

37. Are there convenient examples of monotone set valued maps? Yes, there are. Let X be a Banach space and let $\phi : X \to (-\infty, \infty]$ be convex, lower semicontinuous, and proper. See Problem 28 for a discussion of lower semicontinuous. Proper means that $\phi(x) < \infty$ for some x. Convex means the usual thing. $\phi(tx + (1-t)y) \leq t\phi(x) + (1-t)\phi(y)$ where $t \in [0, 1]$. Then $x^* \in \partial\phi(x)$ means that

$$\langle x^*, z - x \rangle \leq \phi(z) - \phi(x), \text{ for all } z \in X$$

Show that if $x^* \in \partial\phi(x)$, then $\phi(x) < \infty$. The set of points x where $\phi(x) < \infty$ is called the domain of ϕ denoted as $D(\phi)$. Also show that if $[x, x^*], [\hat{x}, \hat{x}^*]$ are two points of the graph of $\partial\phi$, then $\langle \hat{x}^* - x^*, \hat{x} - x \rangle \geq 0$ so that $\partial\phi$ is an example of a monotone graph. You might wonder whether this graph is nonempty. See the next problem for a partial answer to this question. Of course the above problem pertains to finite dimensional spaces so you could just take any $\phi : \mathbb{R}^n \to \mathbb{R}$ which is convex and differentiable. You can see that in this case the subgradient coincides with the derivative.

38. Let $\phi : \mathbb{R}^n \to \mathbb{R}$ be convex, proper lower semicontinuous, and bounded below. Show that the graph of $\partial\phi$ is nonempty. **Hint:** Just consider $\psi(\boldsymbol{x}) = |\boldsymbol{x}|^2 + \phi(\boldsymbol{x})$ and observe that this is coercive. Then argue using convexity that $\partial\psi(x) = \partial\phi(x) + 2x$. (You don't need to assume that ϕ is bounded below but it is convenient to assume this.)

39. Suppose $\boldsymbol{f} : [0,T] \times \mathbb{R}^n \to \mathbb{R}^n$ is continuous and an estimate of the following form holds. $(\boldsymbol{f}(t,\boldsymbol{x}),\boldsymbol{x}) \leq A + B|\boldsymbol{x}|^2$ Show that there exists a solution to the initial value problem $\boldsymbol{x}' = \boldsymbol{f}(t,\boldsymbol{x})$, $\boldsymbol{x}(0) = \boldsymbol{x}_0$ for $t \in [0,T]$.

40. In the above problem, suppose that $-\boldsymbol{f} + \alpha I$ is monotone for large enough α in addition to the estimate of that problem. Show that then there is only one solution to the problem. In fact, show that the solution depends continuously on the initial data.

41. It was shown that if $f : X \to X$ is locally Lipschitz where X is a Banach space. Then there exists a unique local solution to the IVP

$$y' = f(y), \ \ y(0) = y_0$$

If f is bounded, then in fact the solutions exists on $[0,T]$ for any $T > 0$. Show that it suffices to assume that $\|f(y)\| \leq a + b\|y\|$.

42. Suppose $f(\cdot,\cdot) : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ is continuous and also that $|f(t,x)| \leq M$ for all (t,x). Show that there exists a solution to the initial value problem

$$x' = f(t,x), \ x(0) = x_0 \in \mathbb{R}^n$$

for $t \in [0,T]$. **Hint:** You might consider $T : C([0,T],\mathbb{R}^n) \to C([0,T],\mathbb{R}^n)$ given by $Fx(t) \equiv x_0 + \int_0^t f(s,x(s))\,ds$. Argue that F has a fixed point using the Schauder fixed point theorem.

43. Remove the assumption that $|f(t,x)| \leq M$ at the expense of obtaining only a local solution.

Hint: You can consider the closed set in $\mathbb{R}^n$ $B = \overline{B(x_0,R)}$ where R is some positive number. Let P be the projection onto B.

44. In the Schauder fixed point theorem, eliminate the assumption that K is closed. **Hint:** You can argue that the $\{y_i\}$ in the approximation can be in $f(K)$.

45. Show that there is no one to one continuous function

$$\boldsymbol{f} : [0,1] \to \left\{(x,y) : x^2 + y^2 \leq 1\right\}$$

such that $\boldsymbol{f}$ is onto.

Chapter 7

The Derivative

7.1 Limits of a Function

As in the case of scalar valued functions of one variable, a concept closely related to continuity is that of the **limit of a function**. The notion of limit of a function makes sense at points $\boldsymbol{x}$, which are limit points of $D(\boldsymbol{f})$ and this concept is defined next. In all that follows $(V,\|\cdot\|)$ and $(W,\|\cdot\|)$ are two normed linear spaces. Recall the definition of limit point first.

Definition 7.1.1 *Let $A \subseteq W$ be a set. A point $\boldsymbol{x}$, is a limit point of A if $B(\boldsymbol{x},r)$ contains infinitely many points of A for every $r > 0$.*

Definition 7.1.2 *Let $\boldsymbol{f} : D(\boldsymbol{f}) \subseteq V \to W$ be a function and let $\boldsymbol{x}$ be a **limit point** of $D(\boldsymbol{f})$. Then*

$$\lim_{\boldsymbol{y}\to\boldsymbol{x}} \boldsymbol{f}(\boldsymbol{y}) = \boldsymbol{L}$$

if and only if the following condition holds. For all $\varepsilon > 0$ there exists $\delta > 0$ such that if

$$0 < \|\boldsymbol{y}-\boldsymbol{x}\| < \delta, \text{ and } \boldsymbol{y} \in D(\boldsymbol{f})$$

then,

$$\|\boldsymbol{L}-\boldsymbol{f}(\boldsymbol{y})\| < \varepsilon.$$

Theorem 7.1.3 *If $\lim_{\boldsymbol{y}\to\boldsymbol{x}} \boldsymbol{f}(\boldsymbol{y}) = \boldsymbol{L}$ and $\lim_{y\to x} \boldsymbol{f}(\boldsymbol{y}) = \boldsymbol{L}_1$, then $\boldsymbol{L} = \boldsymbol{L}_1$.*

Proof: Let $\varepsilon > 0$ be given. There exists $\delta > 0$ such that if $0 < |\boldsymbol{y}-\boldsymbol{x}| < \delta$ and $\boldsymbol{y} \in D(\boldsymbol{f})$, then $\|\boldsymbol{f}(\boldsymbol{y})-\boldsymbol{L}\| < \varepsilon$, $\|\boldsymbol{f}(\boldsymbol{y})-\boldsymbol{L}_1\| < \varepsilon$. Pick such a $\boldsymbol{y}$. There exists one because $\boldsymbol{x}$ is a limit point of $D(\boldsymbol{f})$. Then $\|\boldsymbol{L}-\boldsymbol{L}_1\| \le \|\boldsymbol{L}-\boldsymbol{f}(\boldsymbol{y})\| + \|\boldsymbol{f}(\boldsymbol{y})-\boldsymbol{L}_1\| < \varepsilon+\varepsilon = 2\varepsilon$. Since $\varepsilon > 0$ was arbitrary, this shows $\boldsymbol{L} = \boldsymbol{L}_1$. ■

One can define what it means for $\lim_{\boldsymbol{y}\to\boldsymbol{x}} f(\boldsymbol{x}) = \pm\infty$. as in the case of real valued functions.

Definition 7.1.4 *If $f(\boldsymbol{x}) \in \mathbb{R}$, $\lim_{\boldsymbol{y}\to\boldsymbol{x}} f(\boldsymbol{x}) = \infty$ if for every number l, there exists $\delta > 0$ such that whenever $\|\boldsymbol{y}-\boldsymbol{x}\| < \delta$ and $\boldsymbol{y} \in D(\boldsymbol{f})$, then $f(\boldsymbol{x}) > l$. Also the assertion that $\lim_{\boldsymbol{y}\to\boldsymbol{x}} f(\boldsymbol{x}) = -\infty$ means that for every number l, there exists $\delta > 0$ such that whenever $\|\boldsymbol{y}-\boldsymbol{x}\| < \delta$ and $\boldsymbol{y} \in D(\boldsymbol{f})$, then $f(\boldsymbol{x}) < l$.*

The following theorem is just like the one variable version of calculus.

Theorem 7.1.5 *Suppose $\boldsymbol{f} : D(\boldsymbol{f}) \subseteq V \to \mathbb{F}^m$. Then for $\boldsymbol{x}$ a limit point of $D(\boldsymbol{f})$,*

$$\lim_{y\to x} \boldsymbol{f}(\boldsymbol{y}) = \boldsymbol{L} \tag{7.1}$$

if and only if

$$\lim_{y\to x} f_k(\boldsymbol{y}) = L_k \tag{7.2}$$

where $\boldsymbol{f}(\boldsymbol{y}) \equiv (f_1(\boldsymbol{y}),\cdots,f_p(\boldsymbol{y}))$ and $\boldsymbol{L} \equiv (L_1,\cdots,L_p)$.

Suppose here that f has values in W, a normed linear space and

$$\lim_{y\to\boldsymbol{x}} f(y) = L, \ \lim_{y\to\boldsymbol{x}} g(y) = K$$

where $K,L \in W$. *Then if* $a, b \in \mathbb{F}$,

$$\lim_{y\to x}(af(y)+bg(y)) = aL+bK, \tag{7.3}$$

If W *is an inner product space,*

$$\lim_{y\to x}(f,g)(y) = (L,K) \tag{7.4}$$

If g *is scalar valued with* $\lim_{y\to x} g(y) = K$,

$$\lim_{y\to x} f(y)\,g(y) = LK. \tag{7.5}$$

Also, if h *is a continuous function defined near* L, *then*

$$\lim_{y\to x} h\circ f(y) = h(L). \tag{7.6}$$

Suppose $\lim_{y\to \boldsymbol{x}} f(y) = L$. *If* $\|f(y)-b\| \le r$ *for all* y *sufficiently close to* $\boldsymbol{x}$, *then* $|L-b| \le r$ *also.*

Proof: Suppose 7.1. Then letting $\varepsilon > 0$ be given there exists $\delta > 0$ such that if $0 < \|y-x\| < \delta$, it follows

$$|f_k(y)-L_k| \le \|\boldsymbol{f}(y)-\boldsymbol{L}\| < \varepsilon$$

which verifies 7.2.

Now suppose 7.2 holds. Then letting $\varepsilon > 0$ be given, there exists δ_k such that if $0 < \|y-x\| < \delta_k$, then $|f_k(y)-L_k| < \varepsilon$. Let $0 < \delta < \min(\delta_1,\cdots,\delta_p)$. Then if $0 < \|y-x\| < \delta$, it follows $\|\boldsymbol{f}(y)-\boldsymbol{L}\|_\infty < \varepsilon$. Any other norm on $\mathbb{F}^m$ would work out the same way because the norms are all equivalent.

Each of the remaining assertions follows immediately from the coordinate descriptions of the various expressions and the first part. However, I will give a different argument for these.

The proof of 7.3 is left for you. Now 7.4 is to be verified. Let $\varepsilon > 0$ be given. Then by the triangle inequality,

$$\begin{aligned}|(f,g)(y)-(L,K)| &\le |(f,g)(y)-(f(y),K)| + |(f(y),K)-(L,K)| \\ &\le \|f(y)\|\,\|g(y)-K\| + \|K\|\,\|f(y)-L\|.\end{aligned}$$

There exists δ_1 such that if $0 < \|y-x\| < \delta_1$ and $y \in D(f)$, then $\|f(y)-L\| < 1$,and so for such y, the triangle inequality implies, $\|f(y)\| < 1+\|L\|$. Therefore, for $0 < \|y-x\| < \delta_1$,

$$|(f,g)(y)-(L,K)| \le (1+\|K\|+\|L\|)\,[\|g(y)-K\| + \|f(y)-L\|]. \tag{7.7}$$

Now let $0 < \delta_2$ be such that if $y \in D(f)$ and $0 < \|x-y\| < \delta_2$,

$$\|f(y)-L\| < \frac{\varepsilon}{2(1+\|K\|+\|L\|)},\ \|g(y)-K\| < \frac{\varepsilon}{2(1+\|K\|+\|L\|)}.$$

Then letting $0 < \delta \le \min(\delta_1,\delta_2)$, it follows from 7.7 that $|(f,g)(y)-(L,K)| < \varepsilon$ and this proves 7.4.

The proof of 7.5 is left to you.

Consider 7.6. Since h is continuous near L, it follows that for $\varepsilon > 0$ given, there exists $\eta > 0$ such that if $\|y-L\| < \eta$, then $\|h(y)-h(L)\| < \varepsilon$. Now since $\lim_{y\to x} f(y) = L$, there exists $\delta > 0$ such that if $0 < \|y-x\| < \delta$, then $\|f(y)-L\| < \eta$.Therefore, if $0 < \|y-x\| < \delta$, $\|h(f(y))-h(L)\| < \varepsilon$.

It only remains to verify the last assertion. Assume $\|f(y)-b\| \leq r$. It is required to show that $\|L-b\| \leq r$. If this is not true, then $\|L-b\| > r$. Consider $B(L, \|L-b\|-r)$. Since L is the limit of f, it follows $f(y) \in B(L, \|L-b\|-r)$ whenever $y \in D(f)$ is close enough to x. Thus, by the triangle inequality, $\|f(y)-L\| < \|L-b\|-r$ and so

$$r < \|L-b\| - \|f(y)-L\| \leq |\|b-L\| - \|f(y)-L\|| \leq \|b-f(y)\|,$$

a contradiction to the assumption that $\|b-f(y)\| \leq r$. ■

The relation between continuity and limits is as follows.

Theorem 7.1.6 *For $f : D(f) \to W$ and $x \in D(f)$ a limit point of $D(f)$, f is continuous at x if and only if* $\lim_{y\to x} f(y) = f(x)$.

Proof: First suppose f is continuous at x a limit point of $D(f)$. Then for every $\varepsilon > 0$ there exists $\delta > 0$ such that if $\|x-y\| < \delta$ and $y \in D(f)$, then $|f(x)-f(y)| < \varepsilon$. In particular, this holds if $0 < \|x-y\| < \delta$ and this is just the definition of the limit. Hence $f(x) = \lim_{y\to x} f(y)$.

Next suppose x is a limit point of $D(f)$ and $\lim_{y\to x} f(y) = f(x)$. This means that if $\varepsilon > 0$ there exists $\delta > 0$ such that for $0 < \|x-y\| < \delta$ and $y \in D(f)$, it follows $|f(y)-f(x)| < \varepsilon$. However, if $y = x$, then $|f(y)-f(x)| = |f(x)-f(x)| = 0$ and so whenever $y \in D(f)$ and $\|x-y\| < \delta$, it follows $|f(x)-f(y)| < \varepsilon$, showing f is continuous at x. ■

Example 7.1.7 *Find* $\lim_{(x,y)\to(3,1)} \left(\frac{x^2-9}{x-3}, y\right)$.

It is clear that $\lim_{(x,y)\to(3,1)} \frac{x^2-9}{x-3} = 6$ and $\lim_{(x,y)\to(3,1)} y = 1$. Therefore, this limit equals $(6,1)$.

Example 7.1.8 *Find* $\lim_{(x,y)\to(0,0)} \frac{xy}{x^2+y^2}$.

First of all, observe the domain of the function is $\mathbb{R}^2 \setminus \{(0,0)\}$, every point in $\mathbb{R}^2$ except the origin. Therefore, $(0,0)$ is a limit point of the domain of the function so it might make sense to take a limit. However, just as in the case of a function of one variable, the limit may not exist. In fact, this is the case here. To see this, take points on the line $y = 0$. At these points, the value of the function equals 0. Now consider points on the line $y = x$ where the value of the function equals $1/2$. Since, arbitrarily close to $(0,0)$, there are points where the function equals $1/2$ and points where the function has the value 0, it follows there can be no limit. Just take $\varepsilon = 1/10$ for example. You cannot be within $1/10$ of $1/2$ and also within $1/10$ of 0 at the same time.

Note it is necessary to rely on the definition of the limit much more than in the case of a function of one variable and there are no easy ways to do limit problems for functions of more than one variable. It is what it is and you will not deal with these concepts without suffering and anguish.

7.2 Basic Definitions

The concept of derivative generalizes right away to functions of many variables. However, no attempt will be made to consider derivatives from one side or another. This is because when you consider functions of many variables, there isn't a well defined side. However, it is certainly the case that there are more general notions which include such things. I will present a fairly general notion of the derivative of a function which is defined on a normed vector space which has values in a normed vector space. The case of most interest is that of a function which maps $\mathbb{F}^n$ to $\mathbb{F}^m$ but it is no more trouble to consider the extra generality and it is sometimes useful to have this extra generality because sometimes you want to consider functions defined, for example on subspaces of $\mathbb{F}^n$and it is nice to not have to trouble with ad hoc considerations. Also, you might want to consider $\mathbb{F}^n$ with some norm other than the usual one.

In what follows, X, Y will denote normed vector spaces. Thanks to Theorem 5.2.4 all the definitions and theorems given below work the same for any norm given on the vector spaces.

Let U be an open set in X, and let $\boldsymbol{f} : U \to Y$ be a function.

Definition 7.2.1 *A function $\boldsymbol{g}$ is $\boldsymbol{o}(\boldsymbol{v})$ if*

$$\lim_{\|\boldsymbol{v}\|\to 0} \frac{\boldsymbol{g}(\boldsymbol{v})}{\|\boldsymbol{v}\|} = \boldsymbol{0} \tag{7.8}$$

A function $\boldsymbol{f} : U \to Y$ is differentiable at $\boldsymbol{x} \in U$ if there exists a linear transformation $L \in \mathscr{L}(X,Y)$ such that

$$\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v}) = \boldsymbol{f}(\boldsymbol{x}) + L\boldsymbol{v} + \boldsymbol{o}(\boldsymbol{v})$$

This linear transformation L is the definition of $D\boldsymbol{f}(\boldsymbol{x})$. This derivative is often called the Frechet derivative.

Note that from Theorem 5.2.4 the question whether a given function is differentiable is independent of the norm used on the finite dimensional vector space. That is, a function is differentiable with one norm if and only if it is differentiable with another norm.

The definition 7.8 means the error,

$$\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{x}) - L\boldsymbol{v}$$

converges to $\boldsymbol{0}$ faster than $\|\boldsymbol{v}\|$. Thus the above definition is equivalent to saying

$$\lim_{\|\boldsymbol{v}\|\to 0} \frac{\|\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{x}) - L\boldsymbol{v}\|}{\|\boldsymbol{v}\|} = 0 \tag{7.9}$$

or equivalently,

$$\lim_{\boldsymbol{y}\to\boldsymbol{x}} \frac{\|\boldsymbol{f}(\boldsymbol{y}) - \boldsymbol{f}(\boldsymbol{x}) - D\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x})\|}{\|\boldsymbol{y}-\boldsymbol{x}\|} = 0. \tag{7.10}$$

The symbol, $\boldsymbol{o}(\boldsymbol{v})$ should be thought of as an adjective. Thus, if t and k are constants,

$$\boldsymbol{o}(\boldsymbol{v}) = \boldsymbol{o}(\boldsymbol{v}) + \boldsymbol{o}(\boldsymbol{v}),\ \boldsymbol{o}(t\boldsymbol{v}) = \boldsymbol{o}(\boldsymbol{v}),\ k\boldsymbol{o}(\boldsymbol{v}) = \boldsymbol{o}(\boldsymbol{v})$$

and other similar observations hold.

Theorem 7.2.2 *The derivative is well defined.*

Proof: First note that for a fixed nonzero vector $\boldsymbol{v}$, $\boldsymbol{o}(t\boldsymbol{v}) = \boldsymbol{o}(t)$. This is because

$$\lim_{t\to 0} \frac{\boldsymbol{o}(t\boldsymbol{v})}{|t|} = \lim_{t\to 0} \|\boldsymbol{v}\| \frac{\boldsymbol{o}(t\boldsymbol{v})}{\|t\boldsymbol{v}\|} = \boldsymbol{0}$$

Now suppose both L_1 and L_2 work in the above definition. Then let $\boldsymbol{v}$ be any vector and let t be a real scalar which is chosen small enough that $t\boldsymbol{v}+\boldsymbol{x} \in U$. Then

$$\boldsymbol{f}(\boldsymbol{x}+t\boldsymbol{v}) = \boldsymbol{f}(\boldsymbol{x}) + L_1 t\boldsymbol{v} + \boldsymbol{o}(t\boldsymbol{v}),\ \boldsymbol{f}(\boldsymbol{x}+t\boldsymbol{v}) = \boldsymbol{f}(\boldsymbol{x}) + L_2 t\boldsymbol{v} + \boldsymbol{o}(t\boldsymbol{v}).$$

Therefore, subtracting these two yields $(L_2 - L_1)(t\boldsymbol{v}) = \boldsymbol{o}(t\boldsymbol{v}) = \boldsymbol{o}(t)$. Therefore, dividing by t yields $(L_2 - L_1)(\boldsymbol{v}) = \frac{\boldsymbol{o}(t)}{t}$. Now let $t \to 0$ to conclude that $(L_2 - L_1)(\boldsymbol{v}) = 0$. Since this is true for all $\boldsymbol{v}$, it follows $L_2 = L_1$. This proves the theorem. ■

In the following lemma, $\|D\boldsymbol{f}(\boldsymbol{x})\|$ is the operator norm of the linear transformation, $D\boldsymbol{f}(\boldsymbol{x})$.

Lemma 7.2.3 *Let $\boldsymbol{f}$ be differentiable at $\boldsymbol{x}$. Then $\boldsymbol{f}$ is continuous at $\boldsymbol{x}$ and in fact, there exists $K > 0$ such that whenever $\|\boldsymbol{v}\|$ is small enough,*

$$\|\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{x})\| \leq K\|\boldsymbol{v}\|$$

Also if $\boldsymbol{f}$ is differentiable at $\boldsymbol{x}$, then

$$\boldsymbol{o}(\|\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{x})\|) = \boldsymbol{o}(\boldsymbol{v})$$

Proof: From the definition of the derivative,

$$\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{x}) = D\boldsymbol{f}(\boldsymbol{x})\boldsymbol{v} + \boldsymbol{o}(\boldsymbol{v}).$$

Let $\|\boldsymbol{v}\|$ be small enough that $\frac{\boldsymbol{o}(\|\boldsymbol{v}\|)}{\|\boldsymbol{v}\|} < 1$ so that $\|\boldsymbol{o}(\boldsymbol{v})\| \leq \|\boldsymbol{v}\|$. Then for such $\boldsymbol{v}$,

$$\|\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{x})\| \leq \|D\boldsymbol{f}(\boldsymbol{x})\boldsymbol{v}\| + \|\boldsymbol{v}\| \leq (\|D\boldsymbol{f}(\boldsymbol{x})\| + 1)\|\boldsymbol{v}\|$$

This proves the lemma with $K = \|D\boldsymbol{f}(\boldsymbol{x})\| + 1$. Recall the operator norm discussed in Definition 5.2.2.

The last assertion is implied by the first as follows. Define

$$\boldsymbol{h}(\boldsymbol{v}) \equiv \begin{cases} \frac{\boldsymbol{o}(\|\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v})-\boldsymbol{f}(\boldsymbol{x})\|)}{\|\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v})-\boldsymbol{f}(\boldsymbol{x})\|} \text{ if } \|\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{x})\| \neq 0 \\ \boldsymbol{0} \text{ if } \|\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{x})\| = 0 \end{cases}$$

Then $\lim_{\|\boldsymbol{v}\|\to 0} \boldsymbol{h}(\boldsymbol{v}) = \boldsymbol{0}$ from continuity of $\boldsymbol{f}$ at $\boldsymbol{x}$ which is implied by the first part. Also from the above estimate, if $\|\boldsymbol{v}\|$ is sufficiently small,

$$\left\| \frac{\boldsymbol{o}(\|\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{x})\|)}{\|\boldsymbol{v}\|} \right\| = \|\boldsymbol{h}(\boldsymbol{v})\| \frac{\|\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{x})\|}{\|\boldsymbol{v}\|} \leq \|\boldsymbol{h}(\boldsymbol{v})\| (\|D\boldsymbol{f}(\boldsymbol{x})\| + 1)$$

and $\lim_{\|\boldsymbol{v}\|\to 0} \|\boldsymbol{h}(\boldsymbol{v})\| = 0$. This establishes the second claim. ■

7.3 The Chain Rule

With the above lemma, it is easy to prove the chain rule.

Theorem 7.3.1 *(The chain rule) Let U and V be open sets $U \subseteq X$ and $V \subseteq Y$. Suppose $\boldsymbol{f} : U \to V$ is differentiable at $\boldsymbol{x} \in U$ and suppose $\boldsymbol{g} : V \to \mathbb{F}^q$ is differentiable at $\boldsymbol{f}(\boldsymbol{x}) \in V$. Then $\boldsymbol{g} \circ \boldsymbol{f}$ is differentiable at $\boldsymbol{x}$ and*

$$D(\boldsymbol{g} \circ \boldsymbol{f})(\boldsymbol{x}) = D\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}))\, D\boldsymbol{f}(\boldsymbol{x}).$$

Proof: This follows from a computation. Let $B(\boldsymbol{x},r) \subseteq U$ and let r also be small enough that for $\|\boldsymbol{v}\| \leq r$, it follows that $\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v}) \in V$. Such an r exists because $\boldsymbol{f}$ is continuous at $\boldsymbol{x}$. For $\|\boldsymbol{v}\| < r$, the definition of differentiability of $\boldsymbol{g}$ and $\boldsymbol{f}$ implies

$$\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v})) - \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x})) =$$

$$\begin{aligned}
& D\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}))(\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{x})) + \boldsymbol{o}(\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{x})) \\
= \;& D\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}))[D\boldsymbol{f}(\boldsymbol{x})\boldsymbol{v} + \boldsymbol{o}(\boldsymbol{v})] + \boldsymbol{o}(\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{x})) \\
= \;& D(\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x})))D(\boldsymbol{f}(\boldsymbol{x}))\boldsymbol{v} + \boldsymbol{o}(\boldsymbol{v}) + \boldsymbol{o}(\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{x})) \qquad (7.11) \\
= \;& D(\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x})))D(\boldsymbol{f}(\boldsymbol{x}))\boldsymbol{v} + \boldsymbol{o}(\boldsymbol{v})
\end{aligned}$$

By Lemma 7.2.3. From the definition of the derivative $D(\boldsymbol{g} \circ \boldsymbol{f})(\boldsymbol{x})$ exists and equals $D(\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x})))D(\boldsymbol{f}(\boldsymbol{x}))$. ■

7.4 The Matrix of the Derivative

The case of most interest here is the only one I will discuss. It is the case where $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$, the function being defined on an open subset of $\mathbb{R}^n$. Of course this all generalizes to arbitrary vector spaces and one considers the matrix taken with respect to various bases. However, I am going to restrict to the case just mentioned here. As above, $\boldsymbol{f}$ will be defined and differentiable on an open set $U \subseteq \mathbb{R}^n$.

As discussed in the review material on linear maps, the matrix of $D\boldsymbol{f}(\boldsymbol{x})$ is the matrix having the i^{th} column equal to $D\boldsymbol{f}(\boldsymbol{x})\boldsymbol{e}_i$ and so it is only necessary to compute this. Let t be a small real number such that

$$\frac{\boldsymbol{f}(\boldsymbol{x}+t\boldsymbol{e}_i) - \boldsymbol{f}(\boldsymbol{x}) - D\boldsymbol{f}(\boldsymbol{x})(t\boldsymbol{e}_i)}{t} = \frac{\boldsymbol{o}(t)}{t}$$

Therefore,

$$\frac{\boldsymbol{f}(\boldsymbol{x}+t\boldsymbol{e}_i) - \boldsymbol{f}(\boldsymbol{x})}{t} = D\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{e}_i) + \frac{\boldsymbol{o}(t)}{t}$$

The limit exists on the right and so it exists on the left also. Thus

$$\frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_i} \equiv \lim_{t \to 0} \frac{\boldsymbol{f}(\boldsymbol{x}+t\boldsymbol{e}_i) - \boldsymbol{f}(\boldsymbol{x})}{t} = D\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{e}_i)$$

and so the matrix of the derivative is just the matrix which has the i^{th} column equal to the i^{th} partial derivative of $\boldsymbol{f}$. Note that this shows that whenever $\boldsymbol{f}$ is differentiable, it follows that the partial derivatives all exist. It does not go the other way however as discussed later.

Theorem 7.4.1 *Let $\boldsymbol{f} : U \subseteq \mathbb{F}^n \to \mathbb{F}^m$ and suppose $\boldsymbol{f}$ is differentiable at $\boldsymbol{x}$. Then all the partial derivatives $\frac{\partial f_i(\boldsymbol{x})}{\partial x_j}$ exist and if $J\boldsymbol{f}(\boldsymbol{x})$ is the matrix of the linear transformation, $D\boldsymbol{f}(\boldsymbol{x})$ with respect to the standard basis vectors, then the ij^{th} entry is given by $\frac{\partial f_i}{\partial x_j}(\boldsymbol{x})$ also denoted as $f_{i,j}$ or f_{i,x_j}. It is the matrix whose i^{th} column is*

$$\frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_i} \equiv \lim_{t\to 0} \frac{\boldsymbol{f}(\boldsymbol{x}+t\boldsymbol{e}_i)-\boldsymbol{f}(\boldsymbol{x})}{t}.$$

Of course there is a generalization of this idea called the directional derivative.

Definition 7.4.2 *In general, the symbol $D_{\boldsymbol{v}}\boldsymbol{f}(\boldsymbol{x})$ is defined by*

$$\lim_{t\to 0} \frac{\boldsymbol{f}(\boldsymbol{x}+t\boldsymbol{v})-\boldsymbol{f}(\boldsymbol{x})}{t}$$

where $t \in \mathbb{F}$. In case $|\boldsymbol{v}| = 1, \mathbb{F} = \mathbb{R}$, and the norm is the standard Euclidean norm, this is called the directional derivative. More generally, with no restriction on the size of $\boldsymbol{v}$ and in any linear space, it is called the Gateaux derivative. $\boldsymbol{f}$ is said to be Gateaux differentiable at $\boldsymbol{x}$ if there exists $D_{\boldsymbol{v}}\boldsymbol{f}(\boldsymbol{x})$ such that

$$\lim_{t\to 0} \frac{\boldsymbol{f}(\boldsymbol{x}+t\boldsymbol{v})-\boldsymbol{f}(\boldsymbol{x})}{t} = D_{\boldsymbol{v}}\boldsymbol{f}(\boldsymbol{x})$$

where $\boldsymbol{v} \to D_{\boldsymbol{v}}\boldsymbol{f}(\boldsymbol{x})$ is linear. Thus we say it is Gateaux differentiable if the Gateaux derivative exists for each $\boldsymbol{v}$ and $\boldsymbol{v} \to D_{\boldsymbol{v}}\boldsymbol{f}(\boldsymbol{x})$ is linear. Note that $\frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_i} = D_{\boldsymbol{e}_i}\boldsymbol{f}(\boldsymbol{x})$. [1]

What if all the partial derivatives of $\boldsymbol{f}$ exist? Does it follow that $\boldsymbol{f}$ is differentiable? Consider the following function, $f : \mathbb{R}^2 \to \mathbb{R}$,

$$f(x,y) = \begin{cases} \frac{xy}{x^2+y^2} \text{ if } (x,y) \neq (0,0) \\ 0 \text{ if } (x,y) = (0,0) \end{cases}.$$

Then from the definition of partial derivatives,

$$\lim_{h\to 0} \frac{f(h,0)-f(0,0)}{h} = \lim_{h\to 0} \frac{0-0}{h} = 0$$

and

$$\lim_{h\to 0} \frac{f(0,h)-f(0,0)}{h} = \lim_{h\to 0} \frac{0-0}{h} = 0$$

However f is not even continuous at $(0,0)$ which may be seen by considering the behavior of the function along the line $y = x$ and along the line $x = 0$. By Lemma 7.2.3 this implies f is not differentiable. Therefore, it is necessary to consider the correct definition of the derivative given above if you want to get a notion which generalizes the concept of the derivative of a function of one variable in such a way as to preserve continuity whenever the function is differentiable.

What if the one dimensional derivative in the definition of the Gateaux derivative exists for all nonzero $\boldsymbol{v}$? Is the function differentiable then? Maybe not. See Problem 12 in the exercises for example.

[1] René Gateaux was one of the many young French men killed in world war I. This derivative is named after him, but it developed naturally from ideas used in the calculus of variations which were due to Euler and Lagrange back in the 1700's.

7.5 A Mean Value Inequality

The following theorem will be very useful in much of what follows. It is a version of the mean value theorem as is the next lemma. The mean value theorem depends on the function having values in $\mathbb{R}$ and in the lemma and theorem, it has values in a normed vector space.

Lemma 7.5.1 *Let Y be a normed vector space and suppose $\boldsymbol{h}:[0,1]\to Y$ is continuous and differentiable from the right and satisfies $\left\|\boldsymbol{h}'(t)\right\|\leq M$, $M\geq 0$. Then $\|\boldsymbol{h}(1)-\boldsymbol{h}(0)\|\leq M$.*

Proof: Let $\varepsilon>0$ be given and let

$$S\equiv\{t\in[0,1]:\ \text{for all } s\in[0,t],\|\boldsymbol{h}(s)-\boldsymbol{h}(0)\|\leq(M+\varepsilon)s\}$$

Then $0\in S$. Let $t=\sup S$. Then by continuity of $\boldsymbol{h}$ it follows

$$\|\boldsymbol{h}(t)-\boldsymbol{h}(0)\|=(M+\varepsilon)t \tag{7.12}$$

Suppose $t<1$. Then there exist positive numbers, h_k decreasing to 0 such that

$$\|\boldsymbol{h}(t+h_k)-\boldsymbol{h}(0)\|>(M+\varepsilon)(t+h_k)$$

and now it follows from 7.12 and the triangle inequality that

$$\begin{aligned}&\|\boldsymbol{h}(t+h_k)-\boldsymbol{h}(t)\|+\|\boldsymbol{h}(t)-\boldsymbol{h}(0)\|\\ =\ &\|\boldsymbol{h}(t+h_k)-\boldsymbol{h}(t)\|+(M+\varepsilon)t>(M+\varepsilon)(t+h_k)\end{aligned}$$

Thus

$$\|\boldsymbol{h}(t+h_k)-\boldsymbol{h}(t)\|>(M+\varepsilon)h_k$$

Now dividing by h_k and letting $k\to\infty$, $\left\|\boldsymbol{h}'(t)\right\|\geq M+\varepsilon$,a contradiction. Thus $t=1$. Since ε is arbitrary, the conclusion of the lemma follows. ■

Theorem 7.5.2 *Suppose U is an open subset of X and $\boldsymbol{f}:U\to Y$ has the property that $D\boldsymbol{f}(\boldsymbol{x})$ exists for all $\boldsymbol{x}$ in U and that, $\boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x})\in U$ for all $t\in[0,1]$. (The line segment joining the two points lies in U.) Suppose also that for all points on this line segment,*

$$\|D\boldsymbol{f}(\boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x}))\|\leq M.$$

Then

$$\|\boldsymbol{f}(\boldsymbol{y})-\boldsymbol{f}(\boldsymbol{x})\|\leq M\|\boldsymbol{y}-\boldsymbol{x}\|.$$

Proof: Let $\boldsymbol{h}(t)\equiv\boldsymbol{f}(\boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x}))$. Then by the chain rule,

$$\boldsymbol{h}'(t)=D\boldsymbol{f}(\boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x}))(\boldsymbol{y}-\boldsymbol{x})$$

and so

$$\left\|\boldsymbol{h}'(t)\right\|=\|D\boldsymbol{f}(\boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x}))(\boldsymbol{y}-\boldsymbol{x})\|\leq M\|\boldsymbol{y}-\boldsymbol{x}\|$$

by Lemma 7.5.1, $\|\boldsymbol{h}(1)-\boldsymbol{h}(0)\|=\|\boldsymbol{f}(\boldsymbol{y})-\boldsymbol{f}(\boldsymbol{x})\|\leq M\|\boldsymbol{y}-\boldsymbol{x}\|$.■

7.6 Existence of the Derivative, C^1 Functions

There is a way to get the differentiability of a function from the existence and continuity of one dimensional directional derivatives. The following theorem is the main result. It gives easy to verify one dimensional conditions for the existence of the derivative. It also shows that the Gateaux derivative $D_{\boldsymbol{v}} f(\boldsymbol{x})$ is linear in $\boldsymbol{v}$, a requirement for a function to be Gateaux differentiable, so it shows Gateaux differentiability also. The meaning of $\|\cdot\|$ will be determined by context in what follows. This theorem says that if the Gateaux derivatives exist for each vector in a basis and they are also continuous, then the function is differentiable.

Theorem 7.6.1 *Let X be a normed vector space having basis $\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n\}$ and let Y be another normed vector space. Let U be an open set in X and let $\boldsymbol{f} : U \to Y$ have the property that the one dimensional limits*

$$D_{\boldsymbol{v}_k} \boldsymbol{f}(\boldsymbol{x}) \equiv \lim_{t \to 0} \frac{\boldsymbol{f}(\boldsymbol{x} + t\boldsymbol{v}_k) - \boldsymbol{f}(\boldsymbol{x})}{t}$$

exist and $\boldsymbol{x} \to D_{\boldsymbol{v}_k} \boldsymbol{f}(\boldsymbol{x})$ are continuous functions of $\boldsymbol{x} \in U$ as functions with values in Y. Then $D\boldsymbol{f}(\boldsymbol{x})$ exists and

$$D\boldsymbol{f}(\boldsymbol{x})\, \boldsymbol{v} = \sum_{k=1}^{n} D_{\boldsymbol{v}_k} \boldsymbol{f}(\boldsymbol{x})\, a_k$$

where $\boldsymbol{v} = \sum_{k=1}^{n} a_k \boldsymbol{v}_k$. Furthermore, $\boldsymbol{x} \to D\boldsymbol{f}(\boldsymbol{x})$ is continuous; that is

$$\lim_{\boldsymbol{y} \to \boldsymbol{x}} \|D\boldsymbol{f}(\boldsymbol{y}) - D\boldsymbol{f}(\boldsymbol{x})\| = 0.$$

Proof: Let $\boldsymbol{v} = \sum_{k=1}^{n} a_k \boldsymbol{v}_k$ where all a_k are small enough that for all k,

$$\boldsymbol{x} + \sum_{j=1}^{k} a_j \boldsymbol{v}_j \in U.$$

Let $\sum_{k=1}^{0} \equiv \boldsymbol{0}$. Then

$$\boldsymbol{f}(\boldsymbol{x} + \boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{x}) = \sum_{k=1}^{n} \left[\boldsymbol{f}\left(\boldsymbol{x} + \sum_{j=1}^{k} a_j \boldsymbol{v}_j \right) - \boldsymbol{f}\left(\boldsymbol{x} + \sum_{j=1}^{k-1} a_j \boldsymbol{v}_j \right) \right]$$

$$= \sum_{k=1}^{n} \left(\boldsymbol{f}(\boldsymbol{x} + a_k \boldsymbol{v}_k) - \boldsymbol{f}(\boldsymbol{x}) \right)$$

$$+ \sum_{k=1}^{n} \left[\left(\boldsymbol{f}\left(\boldsymbol{x} + \sum_{j=1}^{k} a_j \boldsymbol{v}_j \right) - \boldsymbol{f}(\boldsymbol{x} + a_k \boldsymbol{v}_k) \right) - \left(\boldsymbol{f}\left(\boldsymbol{x} + \sum_{j=1}^{k-1} a_j \boldsymbol{v}_j \right) - \boldsymbol{f}(\boldsymbol{x}) \right) \right] \tag{7.13}$$

Consider the k^{th} term in 7.13. If $a_k = 0$ the term is $\boldsymbol{0}$ so assume $a_k \neq 0$. Let

$$\boldsymbol{h}(t) \equiv \boldsymbol{f}\left(\boldsymbol{x} + \sum_{j=1}^{k-1} a_j \boldsymbol{v}_j + t a_k \boldsymbol{v}_k \right) - \boldsymbol{f}(\boldsymbol{x} + t a_k \boldsymbol{v}_k)$$

for $t \in [0,1]$. Thus this term is $\boldsymbol{h}(1)-\boldsymbol{h}(0)$. Then

$$\begin{aligned}\boldsymbol{h}'(t) &= a_k \lim_{h\to 0} \frac{1}{a_k h}\left(\boldsymbol{f}\left(\boldsymbol{x}+\sum_{j=1}^{k-1} a_j\boldsymbol{v}_j+(t+h)a_k\boldsymbol{v}_k\right)-\boldsymbol{f}\left(\boldsymbol{x}+(t+h)a_k\boldsymbol{v}_k\right)\right.\\ &\quad \left.-\left(\boldsymbol{f}\left(\boldsymbol{x}+\sum_{j=1}^{k-1} a_j\boldsymbol{v}_j+ta_k\boldsymbol{v}_k\right)-\boldsymbol{f}\left(\boldsymbol{x}+ta_k\boldsymbol{v}_k\right)\right)\right)\end{aligned}$$

and this equals

$$\left(D_{\boldsymbol{v}_k}\boldsymbol{f}\left(\boldsymbol{x}+\sum_{j=1}^{k-1} a_j\boldsymbol{v}_j+ta_k\boldsymbol{v}_k\right)-D_{\boldsymbol{v}_k}\boldsymbol{f}\left(\boldsymbol{x}+ta_k\boldsymbol{v}_k\right)\right)a_k \tag{7.14}$$

Now without loss of generality, it can be assumed that the norm on X is given by

$$\|\boldsymbol{v}\| \equiv \max\left\{|a_k| : \boldsymbol{v}=\sum_{j=1}^{n} a_k\boldsymbol{v}_k\right\}$$

because by Theorem 5.2.4 all norms on X are equivalent. Therefore, from 7.14 and the assumption that the Gateaux derivatives are continuous,

$$\begin{aligned}\left\|\boldsymbol{h}'(t)\right\| &= \left\|\left(D_{\boldsymbol{v}_k}\boldsymbol{f}\left(\boldsymbol{x}+\sum_{j=1}^{k-1} a_j\boldsymbol{v}_j+ta_k\boldsymbol{v}_k\right)-D_{\boldsymbol{v}_k}\boldsymbol{f}\left(\boldsymbol{x}+ta_k\boldsymbol{v}_k\right)\right)a_k\right\|\\ &\leq \varepsilon|a_k| \leq \varepsilon\|\boldsymbol{v}\|\end{aligned}$$

provided $\|\boldsymbol{v}\|$ is sufficiently small. Since ε is arbitrary, it follows from Lemma 7.5.1 the expression in 7.13 is $\boldsymbol{o}(\boldsymbol{v})$ because this expression equals a finite sum of terms of the form $\boldsymbol{h}(1)-\boldsymbol{h}(0)$ where $\left\|\boldsymbol{h}'(t)\right\| \leq \varepsilon\|\boldsymbol{v}\|$ whenever $\|\boldsymbol{v}\|$ is small enough. Thus

$$\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v})-\boldsymbol{f}(\boldsymbol{x}) = \sum_{k=1}^{n}\left[\boldsymbol{f}(\boldsymbol{x}+a_k\boldsymbol{v}_k)-\boldsymbol{f}(\boldsymbol{x})\right]+\boldsymbol{o}(\boldsymbol{v})$$

$$= \sum_{k=1}^{n} D_{\boldsymbol{v}_k}\boldsymbol{f}(\boldsymbol{x})a_k + \sum_{k=1}^{n}\left[\boldsymbol{f}(\boldsymbol{x}+a_k\boldsymbol{v}_k)-\left(\boldsymbol{f}(\boldsymbol{x})+D_{\boldsymbol{v}_k}\boldsymbol{f}(\boldsymbol{x})a_k\right)\right]+\boldsymbol{o}(\boldsymbol{v}).$$

Consider the k^{th} term in the second sum.

$$\boldsymbol{f}(\boldsymbol{x}+a_k\boldsymbol{v}_k)-\left(\boldsymbol{f}(\boldsymbol{x})+D_{\boldsymbol{v}_k}\boldsymbol{f}(\boldsymbol{x})a_k\right) = a_k\left(\frac{\boldsymbol{f}(\boldsymbol{x}+a_k\boldsymbol{v}_k)-\boldsymbol{f}(\boldsymbol{x})}{a_k}-D_{\boldsymbol{v}_k}\boldsymbol{f}(\boldsymbol{x})\right)$$

where the expression in the parentheses converges to 0 as $a_k \to 0$. Thus whenever $\|\boldsymbol{v}\|$ is sufficiently small,

$$\left\|\boldsymbol{f}(\boldsymbol{x}+a_k\boldsymbol{v}_k)-\boldsymbol{f}(\boldsymbol{x})-D_{\boldsymbol{v}_k}\boldsymbol{f}(\boldsymbol{x})a_k\right\| \leq \varepsilon|a_k| \leq \varepsilon\|\boldsymbol{v}\|$$

which shows the second sum

$$\sum_{k=1}^{n}\left[\boldsymbol{f}(\boldsymbol{x}+a_k\boldsymbol{v}_k)-\boldsymbol{f}(\boldsymbol{x})-D_{\boldsymbol{v}_k}\boldsymbol{f}(\boldsymbol{x})a_k\right]$$

is $\boldsymbol{o}(\boldsymbol{v})$. Therefore,

$$\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v})-\boldsymbol{f}(\boldsymbol{x})=\sum_{k=1}^{n}D_{\boldsymbol{v}_k}\boldsymbol{f}(\boldsymbol{x})\,a_k+\boldsymbol{o}(\boldsymbol{v}).$$

Defining $D\boldsymbol{f}(\boldsymbol{x})\,\boldsymbol{v}\equiv\sum_{k=1}^{n}D_{\boldsymbol{v}_k}\boldsymbol{f}(\boldsymbol{x})\,a_k$ where $\boldsymbol{v}=\sum_k a_k\boldsymbol{v}_k$, it follows $D\boldsymbol{f}(\boldsymbol{x})\in\mathscr{L}(X,Y)$ and is given by the above formula.

It remains to verify $\boldsymbol{x}\to D\boldsymbol{f}(\boldsymbol{x})$ is continuous.

$$\begin{aligned}\|(D\boldsymbol{f}(\boldsymbol{x})-D\boldsymbol{f}(\boldsymbol{y}))\,\boldsymbol{v}\| &\leq \sum_{k=1}^{n}\left\|\left(D_{\boldsymbol{v}_k}\boldsymbol{f}(\boldsymbol{x})-D_{\boldsymbol{v}_k}\boldsymbol{f}(\boldsymbol{y})\right)a_k\right\|\\ &\leq \max\{|a_k|,k=1,\cdots,n\}\sum_{k=1}^{n}\left\|D_{\boldsymbol{v}_k}\boldsymbol{f}(\boldsymbol{x})-D_{\boldsymbol{v}_k}\boldsymbol{f}(\boldsymbol{y})\right\|\\ &= \|\boldsymbol{v}\|\sum_{k=1}^{n}\left\|D_{\boldsymbol{v}_k}\boldsymbol{f}(\boldsymbol{x})-D_{\boldsymbol{v}_k}\boldsymbol{f}(\boldsymbol{y})\right\|\end{aligned}$$

and so

$$\|D\boldsymbol{f}(\boldsymbol{x})-D\boldsymbol{f}(\boldsymbol{y})\|\leq\sum_{k=1}^{n}\left\|D_{\boldsymbol{v}_k}\boldsymbol{f}(\boldsymbol{x})-D_{\boldsymbol{v}_k}\boldsymbol{f}(\boldsymbol{y})\right\|$$

which proves the continuity of $D\boldsymbol{f}$ because of the assumption $\boldsymbol{x}\to D_{\boldsymbol{v}_k}\boldsymbol{f}(\boldsymbol{x})$ are continuous as functions into Y. ■

This motivates the following definition of what it means for a function to be C^1.

Definition 7.6.2 *Let U be an open subset of a normed finite dimensional vector space, X and let $\boldsymbol{f}:U\to Y$ another finite dimensional normed vector space. Then $\boldsymbol{f}$ is said to be C^1 if there exists a basis for $X,\{\boldsymbol{v}_1,\cdots,\boldsymbol{v}_n\}$ such that the Gateaux derivatives,$D_{\boldsymbol{v}_k}\boldsymbol{f}(\boldsymbol{x})$ exist on U and are continuous functions of $\boldsymbol{x}$.*

Note that as a special case where $X=\mathbb{R}^n$, you could let the $\boldsymbol{v}_k=\boldsymbol{e}_k$ and the condition would reduce to nothing more than a statement that the partial derivatives $\frac{\partial\boldsymbol{f}}{\partial x_i}$ are all continuous. If $X=\mathbb{R}$, this is not a very interesting condition. It would say the derivative exists if the derivative exists and is continuous.

Here is another definition of what it means for a function to be C^1.

Definition 7.6.3 *Let U be an open subset of a normed finite dimensional vector space, X and let $\boldsymbol{f}:U\to Y$ another finite dimensional normed vector space. Then $\boldsymbol{f}$ is said to be C^1 if $\boldsymbol{f}$ is differentiable and $\boldsymbol{x}\to D\boldsymbol{f}(\boldsymbol{x})$ is continuous as a map from U to $\mathscr{L}(X,Y)$.*

Now the following major theorem states these two definitions are equivalent. This is obviously so in the special case where $X=\mathbb{R}^n$ and the special basis is the usual one because, as observed earlier, the matrix of $D\boldsymbol{f}(\boldsymbol{x})$ is just the one which has for its columns the partial derivatives which are given to be continuous.

Theorem 7.6.4 *Let U be an open subset of a normed finite dimensional vector space X and let $\boldsymbol{f}:U\to Y$ another finite dimensional normed vector space. Then the two definitions above are equivalent.*

Proof: It was shown in Theorem 7.6.1, the one about the continuity of the Gateaux derivatives yielding differentiability, that Definition 7.6.2 implies 7.6.3. Suppose then that Definition 7.6.3 holds. Then if $\boldsymbol{v}$ is any vector,

$$\begin{aligned}\lim_{t\to 0}\frac{\boldsymbol{f}(\boldsymbol{x}+t\boldsymbol{v})-\boldsymbol{f}(\boldsymbol{x})}{t} &= \lim_{t\to 0}\frac{D\boldsymbol{f}(\boldsymbol{x})t\boldsymbol{v}+\boldsymbol{o}(t\boldsymbol{v})}{t}\\ &= D\boldsymbol{f}(\boldsymbol{x})\boldsymbol{v}+\lim_{t\to 0}\frac{\boldsymbol{o}(t\boldsymbol{v})}{t}=D\boldsymbol{f}(\boldsymbol{x})\boldsymbol{v}\end{aligned}$$

Thus $D_{\boldsymbol{v}}\boldsymbol{f}(\boldsymbol{x})$ exists and equals $D\boldsymbol{f}(\boldsymbol{x})\boldsymbol{v}$. By continuity of $\boldsymbol{x}\to D\boldsymbol{f}(\boldsymbol{x})$, this establishes continuity of $\boldsymbol{x}\to D_{\boldsymbol{v}}\boldsymbol{f}(\boldsymbol{x})$ and proves the theorem. ■

Note that the proof of the theorem also implies the following corollary.

Corollary 7.6.5 *Let U be an open subset of a normed finite dimensional vector space, X and let $\boldsymbol{f}:U\to Y$ another finite dimensional normed vector space. Then if there is a basis of $X,\{\boldsymbol{v}_1,\cdots,\boldsymbol{v}_n\}$ such that the Gateaux derivatives, $D_{\boldsymbol{v}_k}\boldsymbol{f}(\boldsymbol{x})$ exist and are continuous. Then all Gateaux derivatives, $D_{\boldsymbol{v}}\boldsymbol{f}(\boldsymbol{x})$ exist and are continuous for all $\boldsymbol{v}\in X$.*

From now on, whichever definition is more convenient will be used.

7.7 Higher Order Derivatives

If $f:U\subseteq X\to Y$ for U an open set, then $\boldsymbol{x}\to D\boldsymbol{f}(\boldsymbol{x})$ is a mapping from U to $\mathscr{L}(X,Y)$, a normed vector space. Therefore, it makes perfect sense to ask whether this function is also differentiable.

Definition 7.7.1 *The following is the definition of the second derivative.*

$$D^2\boldsymbol{f}(\boldsymbol{x})\equiv D\left(D\boldsymbol{f}(\boldsymbol{x})\right).$$

Thus, $D\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v})-D\boldsymbol{f}(\boldsymbol{x})=D^2\boldsymbol{f}(\boldsymbol{x})\boldsymbol{v}+\boldsymbol{o}(\boldsymbol{v})$.This implies

$$D^2\boldsymbol{f}(\boldsymbol{x})\in\mathscr{L}(X,\mathscr{L}(X,Y)),\ D^2\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{u})(\boldsymbol{v})\in Y,$$

and the map $(\boldsymbol{u},\boldsymbol{v})\to D^2\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{u})(\boldsymbol{v})$ is a bilinear map having values in Y. In other words, the two functions,

$$\boldsymbol{u}\to D^2\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{u})(\boldsymbol{v}),\ \boldsymbol{v}\to D^2\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{u})(\boldsymbol{v})$$

are both linear.

The same pattern applies to taking higher order derivatives. Thus,

$$D^3\boldsymbol{f}(\boldsymbol{x})\equiv D\left(D^2\boldsymbol{f}(\boldsymbol{x})\right)$$

and $D^3\boldsymbol{f}(\boldsymbol{x})$ may be considered as a trilinear map having values in Y. In general $D^k\boldsymbol{f}(\boldsymbol{x})$ may be considered a k linear map. This means $(\boldsymbol{u}_1,\cdots,\boldsymbol{u}_k)\to D^k\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{u}_1)\cdots(\boldsymbol{u}_k)$ has the property $\boldsymbol{u}_j\to D^k\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{u}_1)\cdots(\boldsymbol{u}_j)\cdots(\boldsymbol{u}_k)$ is linear.

Also, instead of writing $D^2\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{u})(\boldsymbol{v})$, or $D^3\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{u})(\boldsymbol{v})(\boldsymbol{w})$ the following notation is often used.

$$D^2\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{u},\boldsymbol{v})\ \text{ or }D^3\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{u},\boldsymbol{v},\boldsymbol{w})$$

with similar conventions for higher derivatives than 3. Another convention which is often used is the notation $D^k\boldsymbol{f}(\boldsymbol{x})\boldsymbol{v}^k$ instead of $D^k\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{v},\cdots,\boldsymbol{v})$.

Note that for every k, $D^k\boldsymbol{f}$ maps U to a normed vector space. As mentioned above, $D\boldsymbol{f}(\boldsymbol{x})$ has values in $\mathscr{L}(X,Y)$, $D^2\boldsymbol{f}(\boldsymbol{x})$ has values in $\mathscr{L}(X,\mathscr{L}(X,Y))$, etc. Thus it makes sense to consider whether $D^k\boldsymbol{f}$ is continuous. This is described in the following definition.

Definition 7.7.2 *Let U be an open subset of X, a normed vector space, and let $\boldsymbol{f} : U \to Y$. Then $\boldsymbol{f}$ is $C^k(U)$ if $\boldsymbol{f}$ and its first k derivatives are all continuous. Also, $D^k \boldsymbol{f}(\boldsymbol{x})$ when it exists can be considered a Y valued multi-linear function. Sometimes these are called tensors in case $\boldsymbol{f}$ has scalar values.*

7.8 Some Standard Notation

In the case where $X = \mathbb{R}^n$ there is a special notation which is often used to describe higher order mixed partial derivatives. It is called multi-index notation.

Definition 7.8.1 *$\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)$ for $\alpha_1 \cdots \alpha_n$ positive integers is called a multi-index, as before with polynomials. For $\boldsymbol{\alpha}$ a multi-index, $|\boldsymbol{\alpha}| \equiv \alpha_1 + \cdots + \alpha_n$, and if $\boldsymbol{x} \in X$,*

$$\boldsymbol{x} = (x_1, \cdots, x_n),$$

and $\boldsymbol{f}$ a function, define

$$\boldsymbol{x}^{\boldsymbol{\alpha}} \equiv x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n},\ D^{\boldsymbol{\alpha}} \boldsymbol{f}(\boldsymbol{x}) \equiv \frac{\partial^{|\boldsymbol{\alpha}|} \boldsymbol{f}(\boldsymbol{x})}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \cdots \partial x_n^{\alpha_n}}.$$

Then in this special case, the following is another description of what is meant by a C^k function.

Definition 7.8.2 *Let U be an open subset of $\mathbb{R}^n$ and let $\boldsymbol{f} : U \to Y$. Then for k a nonnegative integer, a differentiable function $\boldsymbol{f}$ is C^k if for every $|\boldsymbol{\alpha}| \leq k$, $D^{\boldsymbol{\alpha}} \boldsymbol{f}$ exists and is continuous.*

Theorem 7.8.3 *Let U be an open subset of $\mathbb{R}^n$ and let $\boldsymbol{f} : U \to Y$. Then if $D^r \boldsymbol{f}(\boldsymbol{x})$ exists for $r \leq k$, then $D^r \boldsymbol{f}$ is continuous at $\boldsymbol{x}$ for $r \leq k$ if and only if $D^{\boldsymbol{\alpha}} \boldsymbol{f}$ is continuous at x for each $|\boldsymbol{\alpha}| \leq k$.*

Proof: First consider the case of a single derivative. Then as shown above, the matrix of $D\boldsymbol{f}(\boldsymbol{x})$ is just

$$J(\boldsymbol{x}) \equiv \begin{pmatrix} \frac{\partial \boldsymbol{f}}{\partial x_1}(\boldsymbol{x}) & \cdots & \frac{\partial \boldsymbol{f}}{\partial x_n}(\boldsymbol{x}) \end{pmatrix}$$

and to say that $\boldsymbol{x} \to D\boldsymbol{f}(\boldsymbol{x})$ is continuous is the same as saying that each of these partial derivatives is continuous. Written out in more detail,

$$\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{x}) = D\boldsymbol{f}(\boldsymbol{x})\boldsymbol{v} + \boldsymbol{o}(\boldsymbol{v}) = \sum_{k=1}^{n} \frac{\partial \boldsymbol{f}}{\partial x_k}(\boldsymbol{x}) v_k + \boldsymbol{o}(\boldsymbol{v})$$

Thus $D\boldsymbol{f}(\boldsymbol{x})\boldsymbol{v} = \sum_{k=1}^{n} \frac{\partial \boldsymbol{f}}{\partial x_k}(\boldsymbol{x}) v_k$. Now consider the second derivative.

$$D^2 \boldsymbol{f}(\boldsymbol{x})(\boldsymbol{w})(\boldsymbol{v}) =$$

$$D\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{w})\boldsymbol{v} - D\boldsymbol{f}(\boldsymbol{x})\boldsymbol{v} + \boldsymbol{o}(\boldsymbol{w})(\boldsymbol{v}) = \sum_{k=1}^{n} \left(\frac{\partial \boldsymbol{f}}{\partial x_k}(\boldsymbol{x}+\boldsymbol{w}) - \frac{\partial \boldsymbol{f}}{\partial x_k}(\boldsymbol{x}) \right) v_k + \boldsymbol{o}(\boldsymbol{w})(\boldsymbol{v})$$

$$= \sum_{k=1}^{n} \left(\sum_{j=1}^{n} \frac{\partial^2 \boldsymbol{f}(\boldsymbol{x})}{\partial x_j \partial x_k} w_j + \boldsymbol{o}(\boldsymbol{w}) \right) v_k + \boldsymbol{o}(\boldsymbol{w})(\boldsymbol{v}) = \sum_{j,k} \frac{\partial^2 \boldsymbol{f}(\boldsymbol{x})}{\partial x_j \partial x_k} w_j v_k + \boldsymbol{o}(\boldsymbol{w})(\boldsymbol{v})$$

and so $D^2\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{w})(\boldsymbol{v}) = \sum_{j,k} \frac{\partial^2 \boldsymbol{f}(\boldsymbol{x})}{\partial x_j \partial x_k} w_j v_k$. Hence $D^2\boldsymbol{f}$ is continuous if and only if each of these coefficients $\boldsymbol{x} \to \frac{\partial^2 \boldsymbol{f}(\boldsymbol{x})}{\partial x_j \partial x_k}$ is continuous. Obviously you can continue doing this and conclude that $D^k\boldsymbol{f}$ is continuous if and only if all of the partial derivatives of order up to k are continuous. ■

In practice, this is usually what people are thinking when they say that $\boldsymbol{f}$ is C^k. But as just argued, this is the same as saying that the r linear form $\boldsymbol{x} \to D^r\boldsymbol{f}(\boldsymbol{x})$ is continuous into the appropriate space of linear transformations for each $r \leq k$.

Of course the above is based on the assumption that the first k derivatives exist and gives two equivalent formulations which state that these derivatives are continuous. Can anything be said about the existence of the derivatives based on the existence and continuity of the partial derivatives? As pointed out, if the partial derivatives exist and are continuous, then the function is differentiable and has continuous derivative. However, I want to emphasize the idea of the Cartesian product.

7.9 The Derivative and the Cartesian Product

There are theorems which can be used to get differentiability of a function based on existence and continuity of the partial derivatives. A generalization of this was given above. Here a function defined on a product space is considered. It is very much like what was presented above and could be obtained as a special case but to reinforce the ideas, I will do it from scratch because certain aspects of it are important in the statement of the implicit function theorem.

The following is an important abstract generalization of the concept of partial derivative presented above. Insead of taking the derivative with respect to one variable, it is taken with respect to several but not with respect to others. This vague notion is made precise in the following definition. First here is a lemma.

Lemma 7.9.1 *Suppose U is an open set in $X \times Y$. Then the set, $U_{\boldsymbol{y}}$ defined by*

$$U_{\boldsymbol{y}} \equiv \{\boldsymbol{x} \in X : (\boldsymbol{x}, \boldsymbol{y}) \in U\}$$

is an open set in X. Here $X \times Y$ is a finite dimensional vector space in which the vector space operations are defined componentwise. Thus for $a, b \in \mathbb{F}$,

$$a(\boldsymbol{x}_1, \boldsymbol{y}_1) + b(\boldsymbol{x}_2, \boldsymbol{y}_2) = (a\boldsymbol{x}_1 + b\boldsymbol{x}_2, a\boldsymbol{y}_1 + b\boldsymbol{y}_2)$$

and the norm can be taken to be

$$\|(\boldsymbol{x}, \boldsymbol{y})\| \equiv \max(\|\boldsymbol{x}\|, \|\boldsymbol{y}\|)$$

Proof: Recall by Theorem 5.2.4 it does not matter how this norm is defined and the definition above is convenient. It obviously satisfies most axioms of a norm. The only one which is not obvious is the triangle inequality. I will show this now.

$$\begin{aligned} \|(\boldsymbol{x}, \boldsymbol{y}) + (\boldsymbol{x}_1, \boldsymbol{y}_1)\| &\equiv \|(\boldsymbol{x} + \boldsymbol{x}_1, \boldsymbol{y} + \boldsymbol{y}_1)\| \equiv \max(\|\boldsymbol{x} + \boldsymbol{x}_1\|, \|\boldsymbol{y} + \boldsymbol{y}_1\|) \\ &\leq \max(\|\boldsymbol{x}\| + \|\boldsymbol{x}_1\|, \|\boldsymbol{y}\| + \|\boldsymbol{y}_1\|) \\ &\leq \max(\|\boldsymbol{x}\|, \|\boldsymbol{y}\|) + \max(\|\boldsymbol{x}_1\|, \|\boldsymbol{y}_1\|) \\ &\equiv \|(\boldsymbol{x}, \boldsymbol{y})\| + \|(\boldsymbol{x}_1, \boldsymbol{y}_1)\| \end{aligned}$$

Let $\boldsymbol{x} \in U_{\boldsymbol{y}}$. Then $(\boldsymbol{x},\boldsymbol{y}) \in U$ and so there exists $r > 0$ such that $B((\boldsymbol{x},\boldsymbol{y}),r) \in U$. This says that if $(\boldsymbol{u},\boldsymbol{v}) \in X \times Y$ such that $\|(\boldsymbol{u},\boldsymbol{v}) - (\boldsymbol{x},\boldsymbol{y})\| < r$, then $(\boldsymbol{u},\boldsymbol{v}) \in U$. Thus if

$$\|(\boldsymbol{u},\boldsymbol{y}) - (\boldsymbol{x},\boldsymbol{y})\| = \|\boldsymbol{u} - \boldsymbol{x}\|_X < r,$$

then $(\boldsymbol{u},\boldsymbol{y}) \in U$. This has just said that $B(\boldsymbol{x},r)_X$, the ball taken in X is contained in $U_{\boldsymbol{y}}$. This proves the lemma. ■

Or course one could also consider $U_{\boldsymbol{x}} \equiv \{\boldsymbol{y} : (\boldsymbol{x},\boldsymbol{y}) \in U\}$ in the same way and conclude this set is open in Y. Also, the generalization to many factors yields the same conclusion. In this case, for $\boldsymbol{x} \in \prod_{i=1}^n X_i$, let

$$\|\boldsymbol{x}\| \equiv \max\left(\|\boldsymbol{x}_i\|_{X_i} : \boldsymbol{x} = (\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n)\right)$$

Then a similar argument to the above shows this is a norm on $\prod_{i=1}^n X_i$. Consider the triangle inequality.

$$\|(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n) + (\boldsymbol{y}_1,\cdots,\boldsymbol{y}_n)\| = \max_i\left(\|\boldsymbol{x}_i + \boldsymbol{y}_i\|_{X_i}\right) \leq \max_i\left(\|\boldsymbol{x}_i\|_{X_i} + \|\boldsymbol{y}_i\|_{X_i}\right)$$

$$\leq \max_i\left(\|\boldsymbol{x}_i\|_{X_i}\right) + \max_i\left(\|\boldsymbol{y}_i\|_{X_i}\right) = \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$$

Corollary 7.9.2 *Let $U \subseteq \prod_{i=1}^n X_i$ be an open set and let*

$$U_{(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{i-1},\boldsymbol{x}_{i+1},\cdots,\boldsymbol{x}_n)} \equiv \{\boldsymbol{x} \in \mathbb{F}^{r_i} : (\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{i-1},\boldsymbol{x},\boldsymbol{x}_{i+1},\cdots,\boldsymbol{x}_n) \in U\}.$$

Then $U_{(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{i-1},\boldsymbol{x}_{i+1},\cdots,\boldsymbol{x}_n)}$ is an open set in $\mathbb{F}^{r_i}$.

Proof: Let $\boldsymbol{z} \in U_{(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{i-1},\boldsymbol{x}_{i+1},\cdots,\boldsymbol{x}_n)}$. Then $(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{i-1},\boldsymbol{z},\boldsymbol{x}_{i+1},\cdots,\boldsymbol{x}_n) \equiv \boldsymbol{x} \in U$ by definition. Therefore, since U is open, there exists $r > 0$ such that $B(\boldsymbol{x},r) \subseteq U$. It follows that for $B(\boldsymbol{z},r)_{X_i}$ denoting the ball in X_i, it follows that $B(\boldsymbol{z},r)_{X_i} \subseteq U_{(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{i-1},\boldsymbol{x}_{i+1},\cdots,\boldsymbol{x}_n)}$ because to say that $\|\boldsymbol{z} - \boldsymbol{w}\|_{X_i} < r$ is to say that

$$\|(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{i-1},\boldsymbol{z},\boldsymbol{x}_{i+1},\cdots,\boldsymbol{x}_n) - (\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{i-1},\boldsymbol{w},\boldsymbol{x}_{i+1},\cdots,\boldsymbol{x}_n)\| < r$$

and so $\boldsymbol{w} \in U_{(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{i-1},\boldsymbol{x}_{i+1},\cdots,\boldsymbol{x}_n)}$. ■

Next is a generalization of the partial derivative.

Definition 7.9.3 *Let $\boldsymbol{g} : U \subseteq \prod_{i=1}^n X_i \to Y$, where U is an open set. Then the map*

$$\boldsymbol{z} \to \boldsymbol{g}(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{i-1},\boldsymbol{z},\boldsymbol{x}_{i+1},\cdots,\boldsymbol{x}_n)$$

is a function from the open set in X_i,

$$\{\boldsymbol{z} : \boldsymbol{x} = (\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{i-1},\boldsymbol{z},\boldsymbol{x}_{i+1},\cdots,\boldsymbol{x}_n) \in U\}$$

to Y. When this map is differentiable, its derivative is denoted by $D_i\boldsymbol{g}(\boldsymbol{x})$. To aid in the notation, for $\boldsymbol{v} \in X_i$, let $\boldsymbol{\theta}_i\boldsymbol{v} \in \prod_{i=1}^n X_i$ be the vector $(\boldsymbol{0},\cdots,\boldsymbol{v},\cdots,\boldsymbol{0})$ where the $\boldsymbol{v}$ is in the i^{th} slot and for $\boldsymbol{v} \in \prod_{i=1}^n X_i$, let $\boldsymbol{v}_i$ denote the entry in the i^{th} slot of $\boldsymbol{v}$. Thus, by saying

$$\boldsymbol{z} \to \boldsymbol{g}(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{i-1},\boldsymbol{z},\boldsymbol{x}_{i+1},\cdots,\boldsymbol{x}_n)$$

is differentiable is meant that for $\boldsymbol{v} \in X_i$ sufficiently small,

$$\boldsymbol{g}(\boldsymbol{x} + \boldsymbol{\theta}_i\boldsymbol{v}) - \boldsymbol{g}(\boldsymbol{x}) = D_i\boldsymbol{g}(\boldsymbol{x})\boldsymbol{v} + \boldsymbol{o}(\boldsymbol{v}).$$

Note $D_i\boldsymbol{g}(\boldsymbol{x}) \in \mathscr{L}(X_i,Y)$.

As discussed above, we have the following definition of $C^1(U)$.

Definition 7.9.4 *Let $U \subseteq X$ be an open set. Then $\boldsymbol{f}: U \to Y$ is $C^1(U)$ if $\boldsymbol{f}$ is differentiable and the mapping $\boldsymbol{x} \to D\boldsymbol{f}(\boldsymbol{x})$, is continuous as a function from U to $\mathscr{L}(X,Y)$.*

With this definition of partial derivatives, here is the major theorem. Note the resemblance with the matrix of the derivative of a function having values in $\mathbb{R}^m$ in terms of the partial derivatives.

Theorem 7.9.5 *Let $\boldsymbol{g}, U, \prod_{i=1}^n X_i$, be given as in Definition 7.9.3. Then $\boldsymbol{g}$ is $C^1(U)$ if and only if $D_i\boldsymbol{g}$ exists and is continuous on U for each i. In this case, $\boldsymbol{g}$ is differentiable and*

$$D\boldsymbol{g}(\boldsymbol{x})(\boldsymbol{v}) = \sum_k D_k\boldsymbol{g}(\boldsymbol{x})\boldsymbol{v}_k \tag{7.15}$$

where $\boldsymbol{v} = (\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n)$.

Proof: Suppose then that $D_i\boldsymbol{g}$ exists and is continuous for each i. Note that

$$\sum_{j=1}^k \theta_j \boldsymbol{v}_j = (\boldsymbol{v}_1, \cdots, \boldsymbol{v}_k, \boldsymbol{0}, \cdots, \boldsymbol{0}).$$

Thus $\sum_{j=1}^n \theta_j \boldsymbol{v}_j = \boldsymbol{v}$ and define $\sum_{j=1}^0 \theta_j \boldsymbol{v}_j \equiv \boldsymbol{0}$. Therefore,

$$\boldsymbol{g}(\boldsymbol{x}+\boldsymbol{v}) - \boldsymbol{g}(\boldsymbol{x}) = \sum_{k=1}^n \left[\boldsymbol{g}\left(\boldsymbol{x} + \sum_{j=1}^k \theta_j \boldsymbol{v}_j\right) - \boldsymbol{g}\left(\boldsymbol{x} + \sum_{j=1}^{k-1} \theta_j \boldsymbol{v}_j\right) \right] \tag{7.16}$$

Consider the terms in this sum.

$$\boldsymbol{g}\left(\boldsymbol{x} + \sum_{j=1}^k \theta_j \boldsymbol{v}_j\right) - \boldsymbol{g}\left(\boldsymbol{x} + \sum_{j=1}^{k-1} \theta_j \boldsymbol{v}_j\right) = \boldsymbol{g}(\boldsymbol{x}+\theta_k\boldsymbol{v}_k) - \boldsymbol{g}(\boldsymbol{x}) + \tag{7.17}$$

$$\left(\boldsymbol{g}\left(\boldsymbol{x} + \sum_{j=1}^k \theta_j \boldsymbol{v}_j\right) - \boldsymbol{g}(\boldsymbol{x}+\theta_k\boldsymbol{v}_k)\right) - \left(\boldsymbol{g}\left(\boldsymbol{x} + \sum_{j=1}^{k-1} \theta_j \boldsymbol{v}_j\right) - \boldsymbol{g}(\boldsymbol{x})\right) \tag{7.18}$$

and the expression in 7.18 is of the form $\boldsymbol{h}(\boldsymbol{v}_k) - \boldsymbol{h}(\boldsymbol{0})$ where for small $\boldsymbol{w} \in X_k$,

$$\boldsymbol{h}(\boldsymbol{w}) \equiv \boldsymbol{g}\left(\boldsymbol{x} + \sum_{j=1}^{k-1} \theta_j \boldsymbol{v}_j + \theta_k \boldsymbol{w}\right) - \boldsymbol{g}(\boldsymbol{x} + \theta_k \boldsymbol{w}).$$

Therefore,

$$D\boldsymbol{h}(\boldsymbol{w}) = D_k\boldsymbol{g}\left(\boldsymbol{x} + \sum_{j=1}^{k-1} \theta_j \boldsymbol{v}_j + \theta_k \boldsymbol{w}\right) - D_k\boldsymbol{g}(\boldsymbol{x} + \theta_k \boldsymbol{w})$$

and by continuity, $\|D\boldsymbol{h}(\boldsymbol{w})\| < \varepsilon$ provided $\|\boldsymbol{v}\|$ is small enough. Therefore, by Theorem 7.5.2, the mean value inequality, whenever $\|\boldsymbol{v}\|$ is small enough,

$$\|\boldsymbol{h}(\boldsymbol{v}_k) - \boldsymbol{h}(\boldsymbol{0})\| \leq \varepsilon \|\boldsymbol{v}\|$$

which shows that since ε is arbitrary, the expression in 7.18 is $o(v)$. Now in 7.17,

$$g(x+\theta_k v_k) - g(x) = D_k g(x) v_k + o(v_k) = D_k g(x) v_k + o(v).$$

Therefore, referring to 7.16,

$$g(x+v) - g(x) = \sum_{k=1}^{n} D_k g(x) v_k + o(v)$$

which shows $Dg(x)$ exists and equals the formula given in 7.15. Also $x \to Dg(x)$ is continuous since each of the $D_k g(x)$ are.

Next suppose g is C^1. I need to verify that $D_k g(x)$ exists and is continuous. Let $v \in X_k$ sufficiently small. Then

$$g(x+\theta_k v) - g(x) = Dg(x)\theta_k v + o(\theta_k v) = Dg(x)\theta_k v + o(v)$$

since $\|\theta_k v\| = \|v\|$. Then $D_k g(x)$ exists and equals $Dg(x) \circ \theta_k$. Now $x \to Dg(x)$ is continuous. Since θ_k is linear, it follows from Lemma 5.2.1 that $\theta_k : X_k \to \prod_{i=1}^n X_i$ is also continuous. ■

Note that the above argument also works at a single point x. That is, continuity at x of the partials implies $Dg(x)$ exists and is continuous at x.

The way this is usually used is in the following corollary which has already been obtained. Remember the matrix of $Df(x)$. Recall that if a function is C^1 in the sense that $x \to Df(x)$ is continuous then all the partial derivatives exist and are continuous. The next corollary says that if the partial derivatives do exist and are continuous, then the function is differentiable and has continuous derivative.

Corollary 7.9.6 *Let U be an open subset of $\mathbb{F}^n$ and let $f : U \to \mathbb{F}^m$ be C^1 in the sense that all the partial derivatives of f exist and are continuous. Then f is differentiable and*

$$f(x+v) = f(x) + \sum_{k=1}^{n} \frac{\partial f}{\partial x_k}(x) v_k + o(v).$$

Similarly, if the partial derivatives up to order k exist and are continuous, then the function is C^k in the sense that the first k derivatives exist and are continuous.

7.10 Mixed Partial Derivatives

Continuing with the special case where f is defined on an open set in $\mathbb{F}^n$, I will next consider an interesting result which was known to Euler in around 1734 about mixed partial derivatives. It was proved by Clairaut some time later. It turns out that the mixed partial derivatives, if continuous will end up being equal. Recall the notation $f_x = \frac{\partial f}{\partial x} = D_{e_1} f$ and $f_{xy} = \frac{\partial^2 f}{\partial y \partial x} = D_{e_1 e_2} f$.

Theorem 7.10.1 *Suppose $f : U \subseteq \mathbb{F}^2 \to \mathbb{R}$ where U is an open set on which f_x, f_y, f_{xy} and f_{yx} exist. Then if f_{xy} and f_{yx} are continuous at the point $(x,y) \in U$, it follows*

$$f_{xy}(x,y) = f_{yx}(x,y).$$

Proof: Since U is open, there exists $r>0$ such that $B((x,y),r)\subseteq U$. Now let $|t|,|s|<r/2, t, s$ real numbers and consider

$$\Delta(s,t)\equiv\frac{1}{st}\{\overbrace{f(x+t,y+s)-f(x+t,y)}^{h(t)}-\overbrace{(f(x,y+s)-f(x,y))}^{h(0)}\}. \tag{7.19}$$

Note that $(x+t,y+s)\in U$ because

$$\begin{aligned}|(x+t,y+s)-(x,y)| &= |(t,s)|=\left(t^2+s^2\right)^{1/2}\\ &\leq \left(\frac{r^2}{4}+\frac{r^2}{4}\right)^{1/2}=\frac{r}{\sqrt{2}}<r.\end{aligned}$$

As implied above, $h(t)\equiv f(x+t,y+s)-f(x+t,y)$. Therefore, by the mean value theorem from one variable calculus and the (one variable) chain rule,

$$\begin{aligned}\Delta(s,t) &= \frac{1}{st}(h(t)-h(0))=\frac{1}{st}h'(\alpha t)t\\ &= \frac{1}{s}(f_x(x+\alpha t,y+s)-f_x(x+\alpha t,y))\end{aligned}$$

for some $\alpha\in(0,1)$. Applying the mean value theorem again,

$$\Delta(s,t)=f_{xy}(x+\alpha t,y+\beta s)$$

where $\alpha,\beta\in(0,1)$.

If the terms $f(x+t,y)$ and $f(x,y+s)$ are interchanged in 7.19, $\Delta(s,t)$ is unchanged and the above argument shows there exist $\gamma,\delta\in(0,1)$ such that

$$\Delta(s,t)=f_{yx}(x+\gamma t,y+\delta s).$$

Letting $(s,t)\to(0,0)$ and using the continuity of f_{xy} and f_{yx} at (x,y),

$$\lim_{(s,t)\to(0,0)}\Delta(s,t)=f_{xy}(x,y)=f_{yx}(x,y).\blacksquare$$

The following is obtained from the above by simply fixing all the variables except for the two of interest.

Corollary 7.10.2 *Suppose U is an open subset of X and $f:U\to\mathbb{R}$ has the property that for two indices, k,l, f_{x_k}, f_{x_l}, $f_{x_lx_k}$, and $f_{x_kx_l}$ exist on U and $f_{x_kx_l}$ and $f_{x_lx_k}$ are both continuous at $\boldsymbol{x}\in U$. Then $f_{x_kx_l}(\boldsymbol{x})=f_{x_lx_k}(\boldsymbol{x})$.*

By considering the real and imaginary parts of f in the case where f has values in $\mathbb{C}$ you obtain the following corollary.

Corollary 7.10.3 *Suppose U is an open subset of $\mathbb{F}^n$ and $f:U\to\mathbb{F}$ has the property that for two indices, k,l, f_{x_k}, f_{x_l}, $f_{x_lx_k}$, and $f_{x_kx_l}$ exist on U and $f_{x_kx_l}$ and $f_{x_lx_k}$ are both continuous at $\boldsymbol{x}\in U$. Then $f_{x_kx_l}(\boldsymbol{x})=f_{x_lx_k}(\boldsymbol{x})$.*

Finally, by considering the components of $\boldsymbol{f}$ you get the following generalization.

Corollary 7.10.4 *Suppose U is an open subset of $\mathbb{F}^n$ and $\boldsymbol{f}: U \to \mathbb{F}^m$ has the property that for two indices, k,l, $\boldsymbol{f}_{x_k}$, $\boldsymbol{f}_{x_l}$, $\boldsymbol{f}_{x_l x_k}$, and $\boldsymbol{f}_{x_k x_l}$ exist on U and $\boldsymbol{f}_{x_k x_l}$ and $\boldsymbol{f}_{x_l x_k}$ are both continuous at $\boldsymbol{x} \in U$. Then $\boldsymbol{f}_{x_k x_l}(\boldsymbol{x}) = \boldsymbol{f}_{x_l x_k}(\boldsymbol{x})$.*

This can be generalized to functions which have values in a normed linear space, but I plan to stop with what is given above. One way to proceed would be to reduce to a consideration of the coordinate maps and then apply the above. It would even hold in infinite dimensions through the use of the Hahn Banach theorem. The idea is to reduce to the scalar valued case as above.

In addition, it is obvious that for a function of many variables you could pick any pair and say these are equal if they are both continuous.

It is necessary to assume the mixed partial derivatives are continuous in order to assert they are equal. The following is a well known example [2].

Example 7.10.5 *Let*

$$f(x,y) = \begin{cases} \frac{xy(x^2-y^2)}{x^2+y^2} \text{ if } (x,y) \neq (0,0) \\ 0 \text{ if } (x,y) = (0,0) \end{cases}$$

From the definition of partial derivatives it follows that $f_x(0,0) = f_y(0,0) = 0$. Using the standard rules of differentiation, for $(x,y) \neq (0,0)$,

$$f_x = y\frac{x^4 - y^4 + 4x^2y^2}{(x^2+y^2)^2}, \; f_y = x\frac{x^4 - y^4 - 4x^2y^2}{(x^2+y^2)^2}$$

Now

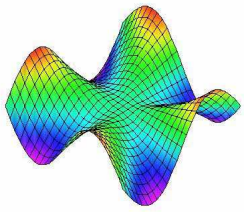
$$f_{xy}(0,0) \equiv \lim_{y\to 0} \frac{f_x(0,y) - f_x(0,0)}{y} = \lim_{y\to 0} \frac{-y^4}{(y^2)^2} = -1$$

while

$$f_{yx}(0,0) \equiv \lim_{x\to 0} \frac{f_y(x,0) - f_y(0,0)}{x} = \lim_{x\to 0} \frac{x^4}{(x^2)^2} = 1$$

showing that although the mixed partial derivatives do exist at $(0,0)$, they are not equal there.

Incidentally, the graph of this function appears very innocent. Its fundamental sickness is not apparent. It is like one of those whited sepulchers mentioned in the Bible.



7.11 A Cofactor Identity

Lemma 7.11.1 *Suppose* $\det(A) = 0$. *Then for all sufficiently small nonzero* ε, *it follows that* $\det(A + \varepsilon I) \neq 0$.

Proof: Let $\det(\lambda I - A) = \lambda^p + a_1\lambda^{p-1} + \cdots + a_{p-1}\lambda + a_p$. First suppose A is a $p \times p$ matrix. If $\det(A) \neq 0$, this will still be true for all ε small enough. Now suppose also that $\det(A) = 0$. Thus, the constant term of $\det(\lambda I - A)$ is 0. Consider $\varepsilon I + A \equiv A_\varepsilon$ for small real ε. The characteristic polynomial of A_ε is

$$\det(\lambda I - A_\varepsilon) = \det((\lambda - \varepsilon) I - A)$$

This is of the form

$$(\lambda - \varepsilon)^p + a_1(\lambda - \varepsilon)^{p-1} + \cdots + (\lambda - \varepsilon)^m a_m$$

where the a_j are the coefficients in the characteristic polynomial for A and $a_k = 0$ for $k > m, a_m \neq 0$. The constant term of this polynomial in λ must be nonzero for all ε small enough because it is of the form

$$(-1)^m \varepsilon^m a_m + (\text{higher order terms in } \varepsilon) = \varepsilon^m [a_m (-1)^m + \varepsilon C(\varepsilon)]$$

which is nonzero for all positive but very small ε. Thus $\varepsilon I + A$ is invertible for all ε small enough but nonzero. ■

Recall that for A an $p \times p$ matrix, $\operatorname{cof}(A)_{ij}$ is the determinant of the matrix which results from deleting the i^{th} row and the j^{th} column and multiplying by $(-1)^{i+j}$. In the proof and in what follows, I am using $D\boldsymbol{g}$ to equal the matrix of the linear transformation $D\boldsymbol{g}$ taken with respect to the usual basis on $\mathbb{R}^p$. Thus $(D\boldsymbol{g})_{ij} = \partial g_i / \partial x_j$ where $\boldsymbol{g} = \sum_i g_i \boldsymbol{e}_i$ for the $\boldsymbol{e}_i$ the standard basis vectors.

Lemma 7.11.2 *Let $\boldsymbol{g} : U \to \mathbb{R}^p$ be C^2 where U is an open subset of $\mathbb{R}^p$. Then*

$$\sum_{j=1}^{p} \operatorname{cof}(D\boldsymbol{g})_{ij,j} = 0,$$

where here $(D\boldsymbol{g})_{ij} \equiv g_{i,j} \equiv \frac{\partial g_i}{\partial x_j}$. Also, $\operatorname{cof}(D\boldsymbol{g})_{ij} = \frac{\partial \det(D\boldsymbol{g})}{\partial g_{i,j}}$.

Proof: From the cofactor expansion theorem,

$$\delta_{kj} \det(D\boldsymbol{g}) = \sum_{i=1}^{p} g_{i,k} \operatorname{cof}(D\boldsymbol{g})_{ij} \tag{7.20}$$

This is because if $k \neq j$, that on the right is the cofactor expansion of a determinant with two equal columns while if $k = j$, it is just the cofactor expansion of the determinant. In particular,

$$\frac{\partial \det(D\boldsymbol{g})}{\partial g_{i,j}} = \operatorname{cof}(D\boldsymbol{g})_{ij} \tag{7.21}$$

which shows the last claim of the lemma. Assume that $D\boldsymbol{g}(\boldsymbol{x})$ is invertible to begin with. Differentiate 7.20 with respect to x_j and sum on j. This yields

$$\sum_{r,s,j} \delta_{kj} \frac{\partial (\det D\boldsymbol{g})}{\partial g_{r,s}} g_{r,sj} = \sum_{ij} g_{i,kj} (\operatorname{cof}(D\boldsymbol{g}))_{ij} + \sum_{ij} g_{i,k} \operatorname{cof}(D\boldsymbol{g})_{ij,j}.$$

Hence, using $\delta_{kj} = 0$ if $j \neq k$ and 7.21,

$$\sum_{rs} (\operatorname{cof}(D\boldsymbol{g}))_{rs} g_{r,sk} = \sum_{rs} g_{r,ks} (\operatorname{cof}(D\boldsymbol{g}))_{rs} + \sum_{ij} g_{i,k} \operatorname{cof}(D\boldsymbol{g})_{ij,j}.$$

Subtracting the first sum on the right from both sides and using the equality of mixed partials,

$$\sum_i g_{i,k}\left(\sum_j (\mathrm{cof}(D\boldsymbol{g}))_{ij,j}\right) = 0.$$

Since it is assumed $D\boldsymbol{g}$ is invertible, this shows $\sum_j (\mathrm{cof}(D\boldsymbol{g}))_{ij,j} = 0$. If $\det(D\boldsymbol{g}) = 0$, use Lemma 7.11.1 to let $\boldsymbol{g}_k(\boldsymbol{x}) = \boldsymbol{g}(\boldsymbol{x}) + \varepsilon_k \boldsymbol{x}$ where $\varepsilon_k \to 0$ and $\det(D\boldsymbol{g} + \varepsilon_k I) \equiv \det(D\boldsymbol{g}_k) \neq 0$. Then

$$\sum_j (\mathrm{cof}(D\boldsymbol{g}))_{ij,j} = \lim_{k\to\infty} \sum_j (\mathrm{cof}(D\boldsymbol{g}_k))_{ij,j} = 0 \blacksquare$$

7.12 Newton's Method

Remember Newton's method from one variable calculus. It was an algorithm for finding the zeros of a function. Beginning with x_k the next iterate was $x_{k+1} = x_k - f'(x_k)^{-1}(f(x_k))$. Of course the same thing can sometimes work in $\mathbb{R}^n$ or even more generally. Here you have a function $\boldsymbol{f}(\boldsymbol{x})$ and you want to locate a zero. Then you could consider the sequence of iterates $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - D\boldsymbol{f}(\boldsymbol{x}_k)^{-1}(\boldsymbol{f}(\boldsymbol{x}_k))$. If the sequence converges to $\boldsymbol{x}$ then you would have $\boldsymbol{x} = \boldsymbol{x} - D\boldsymbol{f}(\boldsymbol{x})^{-1}(\boldsymbol{f}(\boldsymbol{x}))$ and so you would need to have $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}$. In the next section, a modification of this well known method will be used to prove the Implicit function theorem. The modification is that you look for a solution to the equation near $\boldsymbol{x}_0$ and replace the above algorithm with the simpler one $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - D\boldsymbol{f}(\boldsymbol{x}_0)^{-1}(\boldsymbol{f}(\boldsymbol{x}_k))$. Then if $T\boldsymbol{x} = \boldsymbol{x} - D\boldsymbol{f}(\boldsymbol{x}_0)^{-1}(\boldsymbol{f}(\boldsymbol{x}))$, it follows that as long as $\boldsymbol{x}$ is sufficiently close to $\boldsymbol{x}_0, DT(\boldsymbol{x}) = I - D\boldsymbol{f}(\boldsymbol{x}_0)^{-1} D\boldsymbol{f}(\boldsymbol{x})$ and the norm of this transformation is very small so one can use the mean value inequality to conclude that T is a contraction mapping and provide a sequence of iterates which converge to a fixed point. Actually, the situation will be a little more complicated because we will do the implicit function theorem first, but this is the idea.

7.13 Exercises

1. Here are some scalar valued functions of several variables. Determine which of these functions are $o(\boldsymbol{v})$. Here $\boldsymbol{v}$ is a vector in $\mathbb{R}^n$, $\boldsymbol{v} = (v_1, \cdots, v_n)$.

   (a) $v_1 v_2$
   (b) $v_2 \sin(v_1)$
   (c) $v_1^2 + v_2$
   (d) $v_2 \sin(v_1 + v_2)$
   (e) $v_1(v_1 + v_2 + xv_3)$
   (f) $(e^{v_1} - 1 - v_1)$
   (g) $(\boldsymbol{x}\cdot\boldsymbol{v})|\boldsymbol{v}|$

2. Here is a function of two variables. $f(x,y) = x^2 y + x^2$. Find $Df(x,y)$ directly from the definition. Recall this should be a linear transformation which results from multiplication by a 1×2 matrix. Find this matrix.

3. Let $\boldsymbol{f}(x,y) = \begin{pmatrix} x^2 + y \\ y^2 \end{pmatrix}$. Compute the derivative directly from the definition. This should be the linear transformation which results from multiplying by a 2×2 matrix. Find this matrix.

4. You have $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}))$ Here $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}^m$ and $\boldsymbol{g}(\boldsymbol{y}) \in \mathbb{R}^p$. where $\boldsymbol{f}, \boldsymbol{g}$ are appropriately differentiable. Thus $D\boldsymbol{h}(\boldsymbol{x})$ results from multiplication by a matrix. Using the chain rule, give a formula for the ij^{th} entry of this matrix. How does this relate to multiplication of matrices? In other words, you have two matrices which correspond to $D\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}))$ and $D\boldsymbol{f}(\boldsymbol{x})$ Call $\boldsymbol{z} = \boldsymbol{g}(\boldsymbol{y}), \boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$. Then

$$D\boldsymbol{g}(\boldsymbol{y}) = \begin{pmatrix} \frac{\partial \boldsymbol{z}}{\partial y_1} & \cdots & \frac{\partial \boldsymbol{z}}{\partial y_m} \end{pmatrix}, D\boldsymbol{f}(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial \boldsymbol{y}}{\partial x_1} & \cdots & \frac{\partial \boldsymbol{y}}{\partial x_n} \end{pmatrix}$$

Explain the manner in which the ij^{th} entry of $D\boldsymbol{h}(\boldsymbol{x})$ is $\sum_k \frac{\partial z_i}{\partial y_k} \frac{\partial y_y}{\partial x_j}$. This is a review of the way we multiply matrices. what is the i^{th} row of $D\boldsymbol{g}(\boldsymbol{y})$ and the j^{th} column of $D\boldsymbol{f}(\boldsymbol{x})$?

5. Find $f_x, f_y, f_z, f_{xy}, f_{yx}, f_{zy}$ for the following. Verify the mixed partial derivatives are equal.

   (a) $x^2y^3z^4 + \sin(xyz)$

   (b) $\sin(xyz) + x^2yz$

6. Suppose f is a continuous function and $f : U \to \mathbb{R}$ where U is an open set and suppose that $\boldsymbol{x} \in U$ has the property that for all $\boldsymbol{y}$ near $\boldsymbol{x}$, $f(\boldsymbol{x}) \leq f(\boldsymbol{y})$. Prove that if f has all of its partial derivatives at $\boldsymbol{x}$, then $f_{x_i}(\boldsymbol{x}) = 0$ for each x_i. **Hint:** Consider $f(\boldsymbol{x} + t\boldsymbol{v}) = h(t)$. Argue that $h'(0) = 0$ and then see what this implies about $Df(\boldsymbol{x})$.

7. As an important application of Problem 6 consider the following. Experiments are done at n times, $t_1, t_2, \cdots, t_n$ and at each time there results a collection of numerical outcomes. Denote by $\{(t_i, x_i)\}_{i=1}^p$ the set of all such pairs and try to find numbers a and b such that the line $x = at + b$ approximates these ordered pairs as well as possible in the sense that out of all choices of a and b, $\sum_{i=1}^p (at_i + b - x_i)^2$ is as small as possible. In other words, you want to minimize the function of two variables $f(a,b) \equiv \sum_{i=1}^p (at_i + b - x_i)^2$. Find a formula for a and b in terms of the given ordered pairs. You will be finding the formula for the least squares regression line.

8. Let f be a function which has continuous derivatives. Show that $u(t,x) = f(x - ct)$ solves the wave equation $u_{tt} - c^2 \Delta u = 0$. What about $u(x,t) = f(x + ct)$? Here $\Delta u = u_{xx}$.

9. Show that if $\Delta u = \lambda u$ where u is a function of only x, then $e^{\lambda t} u$ solves the heat equation $u_t - \Delta u = 0$. Here $\Delta u = u_{xx}$.

10. Show that if $f(x) = o(x)$, then $f'(0) = 0$.

11. Let $f(x,y)$ be defined on $\mathbb{R}^2$ as follows. $f(x, x^2) = 1$ if $x \neq 0$. Define $f(0,0) = 0$, and $f(x,y) = 0$ if $y \neq x^2$. Show that f is not continuous at $(0,0)$ but that

$$\lim_{h \to 0} \frac{f(ha, hb) - f(0,0)}{h} = 0$$

for (a,b) an arbitrary vector. Thus the Gateaux derivative exists at $(0,0)$ in every direction but f is not even continuous there.

12. Let
$$f(x,y) \equiv \begin{cases} \frac{xy^4}{x^2+y^8} \text{ if } (x,y) \neq (0,0) \\ 0 \text{ if } (x,y) = (0,0) \end{cases}$$
Show that this function is not continuous at $(0,0)$ but that the Gateaux derivative $\lim_{h\to 0} \frac{f(ha,hb)-f(0,0)}{h}$ exists and equals 0 for every vector (a,b).

13. Let U be an open subset of $\mathbb{R}^n$ and suppose that $f : [a,b] \times U \to \mathbb{R}$ satisfies
$$(x,y) \to \frac{\partial f}{\partial y_i}(x,y), (x,y) \to f(x,y)$$
are all continuous. Show that $\int_a^b f(x,y)\,dx$, $\int_a^b \frac{\partial f}{\partial y_i}(x,y)\,dx$ all make sense and that in fact
$$\frac{\partial}{\partial y_i}\left(\int_a^b f(x,y)\,dx\right) = \int_a^b \frac{\partial f}{\partial y_i}(x,y)\,dx$$
Also explain why $y \to \int_a^b \frac{\partial f}{\partial y_i}(x,y)\,dx$ is continuous. **Hint:** You will need to use the theorems from one variable calculus about the existence of the integral for a continuous function. You may also want to use theorems about uniform continuity of continuous functions defined on compact sets.

14. I found this problem in Apostol's book [1]. This is a very important result and is obtained very simply. Read it and fill in any missing details. Let $g(x) \equiv \int_0^1 \frac{e^{-x^2(1+t^2)}}{1+t^2}dt$ and $f(x) \equiv \left(\int_0^x e^{-t^2}dt\right)^2$. Note $\frac{\partial}{\partial x}\left(\frac{e^{-x^2(1+t^2)}}{1+t^2}\right) = -2xe^{-x^2(1+t^2)}$. Explain why this is so. Also show the conditions of Problem 13 are satisfied so that
$$g'(x) = \int_0^1 \left(-2xe^{-x^2(1+t^2)}\right)dt.$$
Now use the chain rule and the fundamental theorem of calculus to find $f'(x)$. Then change the variable in the formula for $f'(x)$ to make it an integral from 0 to 1 and show $f'(x)+g'(x) = 0$. Now this shows $f(x)+g(x)$ is a constant. Show the constant is $\pi/4$ by letting $x \to 0$. Next take a limit as $x \to \infty$ to obtain the following formula for the improper integral, $\int_0^\infty e^{-t^2}dt$, $\left(\int_0^\infty e^{-t^2}dt\right)^2 = \pi/4$. In passing to the limit in the integral for g as $x \to \infty$ you need to justify why that integral converges to 0. To do this, argue the integrand converges uniformly to 0 for $t \in [0,1]$ and then explain why this gives convergence of the integral. Thus $\int_0^\infty e^{-t^2}dt = \sqrt{\pi}/2$.

15. Recall the treatment of integrals of continuous functions in Proposition 5.8.5 or what you used in beginning calculus. The gamma function is defined for $x > 0$ as $\Gamma(x) \equiv \int_0^\infty e^{-t}t^{x-1}dt \equiv \lim_{R\to\infty}\int_0^R e^{-t}t^{x-1}dt$. Show this limit exists. Note you might have to give a meaning to $\int_0^R e^{-t}t^{x-1}dt$ if $x < 1$. Also show that $\Gamma(x+1) = x\Gamma(x)$, $\Gamma(1) = 1$. How does $\Gamma(n)$ for n an integer compare with $(n-1)!$?

16. Show the mean value theorem for integrals. Suppose $f \in C([a,b])$. Then there exists $x \in (a,b)$, not just in $[a,b]$ such that $f(x)(b-a) = \int_a^b f(t)\,dt$. **Hint:** Let $F(x) \equiv \int_a^x f(t)\,dt$ and use the mean value theorem, Theorem 5.8.3 along with $F'(x) = f(x)$.

17. Show, using the Weierstrass approximation theorem that linear combinations of the form $\sum_{i,j} a_{ij} g_i(s) h_j(t)$ where g_i, h_j are continuous functions on $[0,b]$ are dense in $C([0,b]\times[0,b])$, the continuous functions defined on $[0,b]\times[0,b]$ with norm given by
$$\|f\| \equiv \max\{|f(x,y)| : (x,y) \in [0,b]\times[0,b]\}$$
Show that for h,g continuous, $\int_0^b \int_0^s g(s)h(t)\,dtds - \int_0^b \int_t^b g(s)h(t)\,dsdt = 0$. Now explain why if f is in $C([0,b]\times[0,b])$,
$$\int_0^b \int_0^s f(s,t)\,dtds - \int_0^b \int_t^b f(s,t)\,dsdt = 0.$$

18. Zorn's lemma is as follows: You have a nonempty partially ordered set $\mathscr{F}$ with the partial order denoted by $\prec$ and suppose you have the property that every totally ordered subset of $\mathscr{F}$ has an upper bound. Show that it follows that there exists a maximal element $f \in \mathscr{F}$ such that if $f \prec g$ then $f = g$. **Hint:** Use the Hausdorff maximal theorem to show this. In fact, this is equivalent to the Hausdorff maximal theorem.

19. Let $f(x) \equiv \left(\int_0^x e^{-t^2} dt\right)^2$. Use Proposition 5.8.5 which includes the fundamental theorem of calculus and elementary change of variables, show that
$$f'(x) = 2e^{-x^2}\left(\int_0^x e^{-t^2} dt\right) = 2e^{-x^2}\left(\int_0^1 e^{-(xs)^2} x ds\right) = \int_0^1 2xe^{-x^2(1+s^2)} ds.$$
Now use the simple Fubini theorem, Theorem 5.8.6 to write
$$f(x) = \int_0^1 \int_0^x 2te^{-t^2(1+s^2)} dtds.$$
Show $\lim_{x\to\infty} \int_0^x e^{-t^2} dt = \frac{1}{2}\sqrt{\pi}$

Chapter 8

Implicit Function Theorem

8.1 Statement and Proof of the Theorem

Recall the following notation. $\mathscr{L}(X,Y)$ is the space of bounded linear mappings from X to Y where here $(X,\|\cdot\|_X)$ and $(Y,\|\cdot\|_Y)$ are normed linear spaces. Recall that this means that for each $L\in\mathscr{L}(X,Y)$, $\|L\|\equiv\sup_{\|x\|\leq 1}\|Lx\|<\infty$. As shown earlier, this makes $\mathscr{L}(X,Y)$ into a normed linear space. In case X is finite dimensional, $\mathscr{L}(X,Y)$ is the same as the collection of linear maps from X to Y. This was shown earlier. In what follows X,Y will be Banach spaces. If you like, think of them as finite dimensional normed linear spaces, but if you like more generality, just think: complete normed linear space and $\mathscr{L}(X,Y)$ is the space of **bounded** linear maps. In either case, this symbol is given in the following definition.

Definition 8.1.1 *Let $(X,\|\cdot\|_X)$ and $(Y,\|\cdot\|_Y)$ be two normed linear spaces. Then $\mathscr{L}(X,Y)$ denotes the set of linear maps from X to Y which also satisfy the following condition. For $L\in\mathscr{L}(X,Y)$,*

$$\lim_{\|x\|_X\leq 1}\|Lx\|_Y\equiv\|L\|<\infty$$

Recall that this operator norm is less than infinity is always the case where X is finite dimensional. However, if you wish to consider infinite dimensional situations, you assume the operator norm is finite as a qualification for being in $\mathscr{L}(X,Y)$.

Definition 8.1.2 *A complete normed linear space is called a Banach space.*

Theorem 8.1.3 *If Y is a Banach space, then $\mathscr{L}(X,Y)$ is also a Banach space.*

Proof: Let $\{L_n\}$ be a Cauchy sequence in $\mathscr{L}(X,Y)$ and let $x\in X$.

$$\|L_nx-L_mx\|\leq\|x\|\,\|L_n-L_m\|.$$

Thus $\{L_nx\}$ is a Cauchy sequence. Let $Lx=\lim_{n\to\infty}L_nx$. Then, clearly, L is linear because if x_1,x_2 are in X, and a,b are scalars, then

$$\begin{aligned}L(ax_1+bx_2) &= \lim_{n\to\infty}L_n(ax_1+bx_2)=\lim_{n\to\infty}(aL_nx_1+bL_nx_2)\\ &= aLx_1+bLx_2.\end{aligned}$$

Also L is bounded. To see this, note that $\{\|L_n\|\}$ is a Cauchy sequence of real numbers because $|\|L_n\|-\|L_m\||\leq\|L_n-L_m\|$. Hence there exists $K>\sup\{\|L_n\|:n\in\mathbb{N}\}$. Thus, if $x\in X$, $\|Lx\|=\lim_{n\to\infty}\|L_nx\|\leq K\|x\|$. ■

The following theorem is really nice. The series in this theorem is called the Neuman series.

Lemma 8.1.4 *Let $(X,\|\cdot\|)$ is a Banach space, and if $A\in\mathscr{L}(X,X)$ and $\|A\|=r<1$, then $(I-A)^{-1}=\sum_{k=0}^{\infty}A^k\in\mathscr{L}(X,X)$ where the series converges in the Banach space $\mathscr{L}(X,X)$. If O consists of the invertible maps in $\mathscr{L}(X,X)$, then O is open and if $\mathfrak{I}$ is the mapping which takes A to A^{-1}, then $\mathfrak{I}$ is continuous.*

Proof: First of all, why does the series make sense?

$$\left\| \sum_{k=p}^{q} A^k \right\| \leq \sum_{k=p}^{q} \left\| A^k \right\| \leq \sum_{k=p}^{q} \|A\|^k \leq \sum_{k=p}^{\infty} r^k \leq \frac{r^p}{1-r}$$

and so the partial sums are Cauchy in $\mathscr{L}(X,X)$. Therefore, the series converges to something in $\mathscr{L}(X,X)$ by completeness of this normed linear space. Now why is it the inverse?

$$\begin{aligned} \sum_{k=0}^{\infty} A^k (I-A) &\equiv \lim_{n\to\infty} \sum_{k=0}^{n} A^k (I-A) = \lim_{n\to\infty} \left(\sum_{k=0}^{n} A^k - \sum_{k=1}^{n+1} A^k \right) \\ &= \lim_{n\to\infty} \left(I - A^{n+1} \right) = I \end{aligned}$$

because $\left\| A^{n+1} \right\| \leq \|A\|^{n+1} \leq r^{n+1}$. Similarly,

$$(I-A) \sum_{k=0}^{\infty} A^k = \lim_{n\to\infty} \left(I - A^{n+1} \right) = I$$

and so this shows that this series is indeed the desired inverse.

Next suppose $A \in O$ so $A^{-1} \in \mathscr{L}(X,X)$. Then suppose $\|A-B\| < \frac{r}{1+\left\|A^{-1}\right\|}, r < 1$. Does it follow that B is also invertible? $B = A - (A-B) = A\left[I - A^{-1}(A-B) \right]$. Then $\left\| A^{-1}(A-B) \right\| \leq \left\| A^{-1} \right\| \|A-B\| < r$ and so $\left[I - A^{-1}(A-B) \right]^{-1}$ exists. Hence $B^{-1} = \left[I - A^{-1}(A-B) \right]^{-1} A^{-1}$. Thus O is open as claimed. As to continuity, let A, B be as just described. Then using the Neuman series,

$$\|\Im A - \Im B\| = \left\| A^{-1} - \left[I - A^{-1}(A-B) \right]^{-1} A^{-1} \right\|$$

$$\begin{aligned} &= \left\| A^{-1} - \sum_{k=0}^{\infty} \left(A^{-1}(A-B) \right)^k A^{-1} \right\| = \left\| \sum_{k=1}^{\infty} \left(A^{-1}(A-B) \right)^k A^{-1} \right\| \\ &\leq \sum_{k=1}^{\infty} \left\| A^{-1} \right\|^{k+1} \|A-B\|^k = \|A-B\| \left\| A^{-1} \right\|^2 \sum_{k=0}^{\infty} \left\| A^{-1} \right\|^k \left(\frac{r}{1+\left\|A^{-1}\right\|} \right)^k \\ &\leq \|B-A\| \left\| A^{-1} \right\|^2 \frac{1}{1-r}. \end{aligned}$$

Thus $\Im$ is continuous at $A \in O$. ■

Next features the inverse in which there are two different spaces.

Lemma 8.1.5 *Let*

$$O \equiv \{ A \in \mathscr{L}(X,Y) : A^{-1} \in \mathscr{L}(Y,X) \}$$

and let $\Im : O \to \mathscr{L}(Y,X)$, $\Im A \equiv A^{-1}$. *Then* O *is open and* $\Im$ *is in* $C^m(O)$ *for all* $m = 1, 2, \cdots$. *Also*

$$D\Im(A)(B) = -\Im(A)(B)\Im(A). \tag{8.1}$$

In particular, $\Im$ *is continuous.*

Proof: Let $A \in O$ and let $B \in \mathscr{L}(X,Y)$ with $\|B\| \leq \frac{1}{2}\left\|A^{-1}\right\|^{-1}$. Then

$$\left\|A^{-1}B\right\| \leq \left\|A^{-1}\right\| \|B\| \leq \frac{1}{2}$$

So by Lemma 8.1.4,

$$(A+B)^{-1} = \left(I+A^{-1}B\right)^{-1}A^{-1} = \sum_{n=0}^{\infty}(-1)^n\left(A^{-1}B\right)^n A^{-1}$$

$$= \left[I-A^{-1}B+o(B)\right]A^{-1}$$

which shows that O is open and, also,

$$\begin{aligned}\mathfrak{I}(A+B)-\mathfrak{I}(A) &= \sum_{n=0}^{\infty}(-1)^n\left(A^{-1}B\right)^n A^{-1}-A^{-1}\\ &= -A^{-1}BA^{-1}+o(B)\\ &= -\mathfrak{I}(A)(B)\mathfrak{I}(A)+o(B)\end{aligned}$$

which demonstrates 8.1. The reason the left over material is $o(B)$ follows from the observation that $o(B)$ is $\sum_{n=2}^{\infty}(-1)^n\left(A^{-1}B\right)^n A^{-1}$ and so

$$\left\|\sum_{n=2}^{\infty}(-1)^n\left(A^{-1}B\right)^n A^{-1}\right\| \leq \sum_{n=2}^{\infty}\left\|\left(A^{-1}B\right)^n A^{-1}\right\| \leq \left\|A^{-1}\right\|\left\|A^{-1}\right\|^2\|B\|^2\sum_{n=0}^{\infty}\frac{1}{2^n}$$

It follows from this that we can continue taking derivatives of $\mathfrak{I}$. For $\|B_1\|$ small,

$$-\left[D\mathfrak{I}(A+B_1)(B)-D\mathfrak{I}(A)(B)\right] =$$

$$\mathfrak{I}(A+B_1)(B)\mathfrak{I}(A+B_1)-\mathfrak{I}(A)(B)\mathfrak{I}(A)$$

$$\begin{aligned}= \;& \mathfrak{I}(A+B_1)(B)\mathfrak{I}(A+B_1)-\mathfrak{I}(A)(B)\mathfrak{I}(A+B_1)+\\ & \mathfrak{I}(A)(B)\mathfrak{I}(A+B_1)-\mathfrak{I}(A)(B)\mathfrak{I}(A)\\ &= \left[\mathfrak{I}(A)(B_1)\mathfrak{I}(A)+o(B_1)\right](B)\mathfrak{I}(A+B_1)+\\ & \mathfrak{I}(A)(B)\left[\mathfrak{I}(A)(B_1)\mathfrak{I}(A)+o(B_1)\right]\end{aligned}$$

$$\begin{aligned}= \;& \left[\mathfrak{I}(A)(B_1)\mathfrak{I}(A)+o(B_1)\right](B)\left[A^{-1}-A^{-1}B_1A^{-1}+o(B_1)\right]+\\ & \mathfrak{I}(A)(B)\left[\mathfrak{I}(A)(B_1)\mathfrak{I}(A)+o(B_1)\right]\end{aligned}$$

$$= \mathfrak{I}(A)(B_1)\mathfrak{I}(A)(B)\mathfrak{I}(A)+\mathfrak{I}(A)(B)\mathfrak{I}(A)(B_1)\mathfrak{I}(A)+o(B_1)$$

and so

$$D^2\mathfrak{I}(A)(B_1)(B) = \mathfrak{I}(A)(B_1)\mathfrak{I}(A)(B)\mathfrak{I}(A)+\mathfrak{I}(A)(B)\mathfrak{I}(A)(B_1)\mathfrak{I}(A)$$

which shows $\mathfrak{I}$ is $C^2(O)$. Clearly we can continue in this way which shows $\mathfrak{I}$ is in $C^m(O)$ for all $m = 1,2,\cdots$. ■

Here are the two fundamental results presented earlier which will make it easy to prove the implicit function theorem. First is the fundamental mean value inequality.

Theorem 8.1.6 *Suppose U is an open subset of X and $\boldsymbol{f}: U \to Y$ is differentiable on U and $\boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x}) \in U$ for all $t \in [0,1]$. (The line segment joining the two points lies in U.) Suppose also that for all points on this line segment,*

$$\|D\boldsymbol{f}(\boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x}))\| \leq M.$$

Then

$$\|\boldsymbol{f}(\boldsymbol{y})-\boldsymbol{f}(\boldsymbol{x})\| \leq M|\boldsymbol{y}-\boldsymbol{x}|.$$

Next recall the following theorem about fixed points of a contraction map. It was Corollary 3.8.3.

Corollary 8.1.7 *Let B be a closed subset of the complete metric space (X,d) and let $f: B \to X$ be a contraction map*

$$d(f(x),f(\hat{x})) \leq rd(x,\hat{x}),\ r<1.$$

Also suppose ***there exists*** *$x_0 \in B$ such that the sequence of iterates $\{f^n(x_0)\}_{n=1}^{\infty}$ remains in B. Then f has a unique fixed point in B which is the limit of the sequence of iterates. This is a point $x \in B$ such that $f(x)=x$. In the case that $B=\overline{B(x_0,\delta)}$, the sequence of iterates satisfies the inequality*

$$d(f^n(x_0),x_0) \leq \frac{d(x_0,f(x_0))}{1-r}$$

and so it will remain in B if

$$\frac{d(x_0,f(x_0))}{1-r} < \delta.$$

The implicit function theorem deals with the question of solving, $\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})=\boldsymbol{0}$ for $\boldsymbol{x}$ in terms of $\boldsymbol{y}$ and how smooth the solution is. It is one of the most important theorems in mathematics. The proof I will give holds with no change in the context of infinite dimensional complete normed vector spaces when suitable modifications are made on what is meant by $\mathscr{L}(X,Y)$. There are also even more general versions of this theorem than to normed vector spaces.

Recall that for X,Y normed vector spaces, the norm on $X \times Y$ is of the form

$$\|(\boldsymbol{x},\boldsymbol{y})\| = \max(\|\boldsymbol{x}\|,\|\boldsymbol{y}\|).$$

Theorem 8.1.8 *(implicit function theorem) Let X,Y,Z be finite dimensional normed vector spaces and suppose U is an open set in $X \times Y$. Let $\boldsymbol{f}: U \to Z$ be in $C^1(U)$ and suppose*

$$\boldsymbol{f}(\boldsymbol{x}_0,\boldsymbol{y}_0)=\boldsymbol{0},\ D_1\boldsymbol{f}(\boldsymbol{x}_0,\boldsymbol{y}_0)^{-1} \in \mathscr{L}(Z,X). \tag{8.2}$$

Then there exist positive constants, δ,η, such that for every $\boldsymbol{y} \in B(\boldsymbol{y}_0,\eta)$ there exists a unique $\boldsymbol{x}(\boldsymbol{y}) \in B(\boldsymbol{x}_0,\delta)$ such that

$$\boldsymbol{f}(\boldsymbol{x}(\boldsymbol{y}),\boldsymbol{y})=\boldsymbol{0}. \tag{8.3}$$

Furthermore, the mapping, $\boldsymbol{y} \to \boldsymbol{x}(\boldsymbol{y})$ is in $C^1(B(\boldsymbol{y}_0,\eta))$.

Proof: Let $\boldsymbol{T}(\boldsymbol{x},\boldsymbol{y}) \equiv \boldsymbol{x} - D_1\boldsymbol{f}(\boldsymbol{x}_0,\boldsymbol{y}_0)^{-1}\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})$. Therefore, $\boldsymbol{T}(\boldsymbol{x}_0,\boldsymbol{y}_0) = \boldsymbol{x}_0$ and

$$D_1\boldsymbol{T}(\boldsymbol{x},\boldsymbol{y}) = I - D_1\boldsymbol{f}(\boldsymbol{x}_0,\boldsymbol{y}_0)^{-1} D_1\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y}). \tag{8.4}$$

by continuity of the derivative which implies continuity of $D_1\boldsymbol{T}$, it follows there exists $\delta > 0$ such that if $\|\boldsymbol{x}-\boldsymbol{x}_0\| < \delta$ and $\|\boldsymbol{y}-\boldsymbol{y}_0\| < \delta$, then

$$\|D_1\boldsymbol{T}(\boldsymbol{x},\boldsymbol{y})\| < \frac{1}{2},\ D_1\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})^{-1} \text{ exists} \tag{8.5}$$

The second claim follows from Lemma 8.1.5. By the mean value inequality, Theorem 8.1.6, whenever $\boldsymbol{x},\boldsymbol{x}' \in B(\boldsymbol{x}_0,\delta)$ and $\boldsymbol{y} \in B(\boldsymbol{y}_0,\delta)$,

$$\left\|\boldsymbol{T}(\boldsymbol{x},\boldsymbol{y}) - \boldsymbol{T}(\boldsymbol{x}',\boldsymbol{y})\right\| \le \frac{1}{2}\left\|\boldsymbol{x}-\boldsymbol{x}'\right\|. \tag{8.6}$$

Also, it can be assumed δ is small enough that for some M and all such $(\boldsymbol{x},\boldsymbol{y})$,

$$\left\|D_1\boldsymbol{f}(\boldsymbol{x}_0,\boldsymbol{y}_0)^{-1}\right\| \|D_2\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})\| < M \tag{8.7}$$

Next, consider only $\boldsymbol{y}$ such that $\|\boldsymbol{y}-\boldsymbol{y}_0\| < \eta$ where $\eta < \delta$ is so small that

$$\|\boldsymbol{T}(\boldsymbol{x}_0,\boldsymbol{y}) - \boldsymbol{x}_0\| < \frac{\delta}{3}$$

Then for such $\boldsymbol{y}$, consider the mapping $\boldsymbol{T}_{\boldsymbol{y}}(\boldsymbol{x}) = \boldsymbol{T}(\boldsymbol{x},\boldsymbol{y})$. Thus by Corollary 8.1.7, for each $n \in \mathbb{N}$,

$$\delta > \frac{2}{3}\delta \ge \frac{\|\boldsymbol{T}_{\boldsymbol{y}}(\boldsymbol{x}_0) - \boldsymbol{x}_0\|}{1-(1/2)} \ge \left\|\boldsymbol{T}_{\boldsymbol{y}}^n(\boldsymbol{x}_0) - \boldsymbol{x}_0\right\|$$

Then by 8.6, the sequence of iterations of this map $\boldsymbol{T}_{\boldsymbol{y}}$ converges to a unique fixed point $\boldsymbol{x}(\boldsymbol{y})$ in the ball $B(\boldsymbol{x}_0,\delta)$. Thus, from the definition of $\boldsymbol{T}$, $\boldsymbol{f}(\boldsymbol{x}(\boldsymbol{y}),\boldsymbol{y}) = \boldsymbol{0}$. This is the implicitly defined function.

Next we show that this function is Lipschitz continuous. For $\boldsymbol{y},\hat{\boldsymbol{y}}$ in $B(\boldsymbol{y}_0,\eta)$,

$$\begin{aligned}\|\boldsymbol{T}(\boldsymbol{x},\boldsymbol{y}) - \boldsymbol{T}(\boldsymbol{x},\hat{\boldsymbol{y}})\| &= \left\|D_1\boldsymbol{f}(\boldsymbol{x}_0,\boldsymbol{y}_0)^{-1}\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y}) - D_1\boldsymbol{f}(\boldsymbol{x}_0,\boldsymbol{y}_0)^{-1}\boldsymbol{f}(\boldsymbol{x},\hat{\boldsymbol{y}})\right\| \\ &\le M\|\boldsymbol{y}-\hat{\boldsymbol{y}}\|\end{aligned}$$

thanks to the above estimate 8.7 and the mean value inequality, Theorem 8.1.6. Note how convexity of $B(\boldsymbol{y}_0,\eta)$ which says that the line segment joining $\boldsymbol{y},\hat{\boldsymbol{y}}$ is contained in $B(\boldsymbol{y}_0,\eta)$ is important to use this theorem. Then from this,

$$\begin{aligned}\|\boldsymbol{x}(\boldsymbol{y}) - \boldsymbol{x}(\hat{\boldsymbol{y}})\| &= \|\boldsymbol{T}(\boldsymbol{x}(\boldsymbol{y}),\boldsymbol{y}) - \boldsymbol{T}(\boldsymbol{x}(\hat{\boldsymbol{y}}),\hat{\boldsymbol{y}})\| \le \|\boldsymbol{T}(\boldsymbol{x}(\boldsymbol{y}),\boldsymbol{y}) - \boldsymbol{T}(\boldsymbol{x}(\boldsymbol{y}),\hat{\boldsymbol{y}})\| \\ &\quad + \|\boldsymbol{T}(\boldsymbol{x}(\boldsymbol{y}),\hat{\boldsymbol{y}}) - \boldsymbol{T}(\boldsymbol{x}(\hat{\boldsymbol{y}}),\hat{\boldsymbol{y}})\| \\ &\le M\|\boldsymbol{y}-\hat{\boldsymbol{y}}\| + \frac{1}{2}\|\boldsymbol{x}(\boldsymbol{y}) - \boldsymbol{x}(\hat{\boldsymbol{y}})\|\end{aligned}$$

Hence,

$$\|\boldsymbol{x}(\boldsymbol{y}) - \boldsymbol{x}(\hat{\boldsymbol{y}})\| \le 2M\|\boldsymbol{y}-\hat{\boldsymbol{y}}\| \tag{8.8}$$

Finally consider the claim that this implicitly defined function is C^1.

$$\begin{aligned} \mathbf{0} &= \boldsymbol{f}(\boldsymbol{x}(\boldsymbol{y}+\boldsymbol{u}),\boldsymbol{y}+\boldsymbol{u}) - \boldsymbol{f}(\boldsymbol{x}(\boldsymbol{y}),\boldsymbol{y}) \\ &= D_1\boldsymbol{f}(\boldsymbol{x}(\boldsymbol{y}),\boldsymbol{y})(\boldsymbol{x}(\boldsymbol{y}+\boldsymbol{u}) - \boldsymbol{x}(\boldsymbol{y})) + D_2\boldsymbol{f}(\boldsymbol{x}(\boldsymbol{y}),\boldsymbol{y})\boldsymbol{u} \\ &\quad + \boldsymbol{o}(\boldsymbol{x}(\boldsymbol{y}+\boldsymbol{u}) - \boldsymbol{x}(\boldsymbol{y}),\boldsymbol{u}) \end{aligned} \tag{8.9}$$

Consider the last term. $\boldsymbol{o}(\boldsymbol{x}(\boldsymbol{y}+\boldsymbol{u}) - \boldsymbol{x}(\boldsymbol{y}),\boldsymbol{u}) / \|\boldsymbol{u}\|$ equals

$$\begin{cases} \frac{\boldsymbol{o}(\boldsymbol{x}(\boldsymbol{y}+\boldsymbol{u})-\boldsymbol{x}(\boldsymbol{y}),\boldsymbol{u})}{\|(\boldsymbol{x}(\boldsymbol{y}+\boldsymbol{u})-\boldsymbol{x}(\boldsymbol{y}),\boldsymbol{u})\|_{X\times Y}} \frac{\max(\|\boldsymbol{x}(\boldsymbol{y}+\boldsymbol{u})-\boldsymbol{x}(\boldsymbol{y})\|,\|\boldsymbol{u}\|)}{\|\boldsymbol{u}\|} \text{ if } \|(\boldsymbol{x}(\boldsymbol{y}+\boldsymbol{u}) - \boldsymbol{x}(\boldsymbol{y}),\boldsymbol{u})\|_{X\times Y} \neq 0 \\ \mathbf{0} \text{ if } \|(\boldsymbol{x}(\boldsymbol{y}+\boldsymbol{u}) - \boldsymbol{x}(\boldsymbol{y}),\boldsymbol{u})\|_{X\times Y} = 0 \end{cases}$$

Now the Lipschitz condition just established shows that

$$\frac{\max(\|\boldsymbol{x}(\boldsymbol{y}+\boldsymbol{u}) - \boldsymbol{x}(\boldsymbol{y})\|, \|\boldsymbol{u}\|)}{\|\boldsymbol{u}\|}$$

is bounded for nonzero $\boldsymbol{u}$ sufficiently small that $\boldsymbol{y}, \boldsymbol{y}+\boldsymbol{u} \in B(\boldsymbol{y}_0, \eta)$. Therefore,

$$\lim_{\boldsymbol{u}\to\mathbf{0}} \frac{\boldsymbol{o}(\boldsymbol{x}(\boldsymbol{y}+\boldsymbol{u}) - \boldsymbol{x}(\boldsymbol{y}),\boldsymbol{u})}{\|\boldsymbol{u}\|} = \mathbf{0}$$

Then 8.9 shows that

$$\mathbf{0} = D_1\boldsymbol{f}(\boldsymbol{x}(\boldsymbol{y}),\boldsymbol{y})(\boldsymbol{x}(\boldsymbol{y}+\boldsymbol{u}) - \boldsymbol{x}(\boldsymbol{y})) + D_2\boldsymbol{f}(\boldsymbol{x}(\boldsymbol{y}),\boldsymbol{y})\boldsymbol{u} + \boldsymbol{o}(\boldsymbol{u})$$

Therefore, solving for $\boldsymbol{x}(\boldsymbol{y}+\boldsymbol{u}) - \boldsymbol{x}(\boldsymbol{y})$, it follows that

$$\begin{aligned} \boldsymbol{x}(\boldsymbol{y}+\boldsymbol{u}) - \boldsymbol{x}(\boldsymbol{y}) &= -D_1\boldsymbol{f}(\boldsymbol{x}(\boldsymbol{y}),\boldsymbol{y})^{-1} D_2\boldsymbol{f}(\boldsymbol{x}(\boldsymbol{y}),\boldsymbol{y})\boldsymbol{u} + D_1\boldsymbol{f}(\boldsymbol{x}(\boldsymbol{y}),\boldsymbol{y})^{-1}\boldsymbol{o}(\boldsymbol{u}) \\ &= -D_1\boldsymbol{f}(\boldsymbol{x}(\boldsymbol{y}),\boldsymbol{y})^{-1} D_2\boldsymbol{f}(\boldsymbol{x}(\boldsymbol{y}),\boldsymbol{y})\boldsymbol{u} + \boldsymbol{o}(\boldsymbol{u}) \end{aligned}$$

and now, the continuity of the partial derivatives $D_1\boldsymbol{f}$,$D_2\boldsymbol{f}$, continuity of the map $A \to A^{-1}$, along with the continuity of $\boldsymbol{y} \to \boldsymbol{x}(\boldsymbol{y})$ shows that $\boldsymbol{y} \to \boldsymbol{x}(\boldsymbol{y})$ is C^1 with derivative equal to $-D_1\boldsymbol{f}(\boldsymbol{x}(\boldsymbol{y}),\boldsymbol{y})^{-1} D_2\boldsymbol{f}(\boldsymbol{x}(\boldsymbol{y}),\boldsymbol{y})$. ■

The following is a nice result on functional dependence which is an application of the implicit function theorem. See Widder [58].

Example 8.1.9 *Suppose f, g are C^1 near $(x_0, y_0) \in \mathbb{R}^2$ and*
Suppose f, g are C^1 and

1. $\det \begin{pmatrix} f_x(x,y) & f_y(x,y) \\ g_x(x,y) & g_y(x,y) \end{pmatrix} = 0$ *near* (x_0, y_0)
2. $f_x(x_0, y_0) \neq 0$.

Then there is a C^1 function F such that $g(x,y) = F(f(x,y))$ for (x,y) near (x_0, y_0).

Consider $f(x,y) - z = 0$ where $z_0 \equiv f(x_0, y_0)$. By assumption and implicit function theorem, there is a C^1 function $(y,z) \to \phi(y,z)$ so that for (y,z) near (y_0, z_0) the x which solves $f(x,y) - z = 0$ is $x = \phi(y,z)$. In particular, for (x,y) close to (x_0, y_0), $f(\phi(y, f(x,y)), y) - f(x,y) = 0$ and so

$$\phi(y, f(x,y)) = x. \tag{$*$}$$

Also, for (y,z) near (y_0,z_0), it follows $f_x(\phi(y,z),y)\neq 0$.

Since $f(\phi(y,f(x,y)),y)-z=0$, $f_x(\phi(y,z),y)\phi_y(y,z)+f_y(\phi(y,z),y)=0$. It follows from 1. that

$$\begin{aligned}\frac{\partial}{\partial y}g(\phi(y,z),y) &= g_x(\phi(y,z),y)\phi_y(y,z)+g_y(\phi(y,z),y)\\ &= -g_x(\phi(y,z),y)\left(\frac{f_y(\phi(y,z),y)}{f_x(\phi(y,z),y)}\right)+g_y(\phi(y,z),y)\end{aligned}$$

$$=\frac{1}{f_x(\phi(y,z),y)}\begin{pmatrix}-g_x(\phi(y,z),y)f_y(\phi(y,z),y)\\ +g_y(\phi(y,z),y)f_x(\phi(y,z),y)\end{pmatrix}=0$$

Therefore, $g(\phi(y,z),y)$ does not depend on y near $(y_0,z_0)=(y_0,f(x_0,y_0))$. Thus there exists a C^1 function $z\to F(z)$ for z near $f(x_0,y_0)$ such that $g(\phi(y,z),y)=F(z)$. From $*$, for (x,y) near (x_0,y_0),

$$g(x,y)=g(\phi(y,f(x,y)),y)=F(f(x,y))$$

Note that if $g(x,y)=F(f(x,y))$, then $(g_x,g_y)=F'(f(x,y))(f_x,f_y)$ and so the above determinant will equal 0.

The next theorem is a very important special case of the implicit function theorem known as the inverse function theorem. Actually one can also obtain the implicit function theorem from the inverse function theorem. It is done this way in [36], [39] and in [2].

Theorem 8.1.10 *(inverse function theorem) Let $\boldsymbol{x}_0\in U$, an open set in X, and let $\boldsymbol{f}:U\to Y$ where X,Y are finite dimensional normed vector spaces. Suppose*

$$\boldsymbol{f}\text{ is }C^1(U)\ ,\ \text{and}\ D\boldsymbol{f}(\boldsymbol{x}_0)^{-1}\in\mathscr{L}(Y,X). \tag{8.10}$$

Then there exist open sets W, and V such that

$$\boldsymbol{x}_0\in W\subseteq U, \tag{8.11}$$

$$\boldsymbol{f}:W\to V\text{ is one to one and onto}, \tag{8.12}$$

$$\boldsymbol{f}^{-1}\text{ is }C^1, \tag{8.13}$$

Proof: Apply the implicit function theorem to the function $\boldsymbol{F}(\boldsymbol{x},\boldsymbol{y})\equiv\boldsymbol{f}(\boldsymbol{x})-\boldsymbol{y}$ where $\boldsymbol{y}_0\equiv\boldsymbol{f}(\boldsymbol{x}_0)$. Thus the function $\boldsymbol{y}\to\boldsymbol{x}(\boldsymbol{y})$ defined in that theorem is $\boldsymbol{f}^{-1}$. Now let $W\equiv B(\boldsymbol{x}_0,\delta)\cap\boldsymbol{f}^{-1}(B(\boldsymbol{y}_0,\eta))$ and $V\equiv B(\boldsymbol{y}_0,\eta)$.This proves the theorem. ■

8.2 More Derivatives

When you consider a C^k function $\boldsymbol{f}$ defined on an open set U, you obtain the following

$$D\boldsymbol{f}(\boldsymbol{x})\in\mathscr{L}(X,Y),D^2\boldsymbol{f}(\boldsymbol{x})\in\mathscr{L}(X,\mathscr{L}(X,Y)),D^3\boldsymbol{f}(\boldsymbol{x})\in\mathscr{L}(X,\mathscr{L}(X,\mathscr{L}(X,Y)))$$

and so forth. Thus they can each be considered as a linear transformation with values in some vector space. When you consider the vector spaces, you see that these can also be considered as multilinear functions on X with values in Y. Now consider the product of two linear transformations $A(\boldsymbol{y})B(\boldsymbol{y})\boldsymbol{w}$, where everything is given to make sense and here $\boldsymbol{w}$

is an appropriate vector. Then if each of these linear transformations can be differentiated, you would do the following simple computation.

$$\begin{aligned}
&(A(\boldsymbol{y}+\boldsymbol{u})B(\boldsymbol{y}+\boldsymbol{u})-A(\boldsymbol{y})B(\boldsymbol{y}))(\boldsymbol{w}) \\
&= (A(\boldsymbol{y}+\boldsymbol{u})B(\boldsymbol{y}+\boldsymbol{u})-A(\boldsymbol{y})B(\boldsymbol{y}+\boldsymbol{u})+A(\boldsymbol{y})B(\boldsymbol{y}+\boldsymbol{u})-A(\boldsymbol{y})B(\boldsymbol{y}))(\boldsymbol{w}) \\
&= ((DA(\boldsymbol{y})\boldsymbol{u}+\boldsymbol{o}(\boldsymbol{u}))B(\boldsymbol{y}+\boldsymbol{u})+A(\boldsymbol{y})(DB(\boldsymbol{y})\boldsymbol{u}+\boldsymbol{o}(\boldsymbol{u})))(\boldsymbol{w}) \\
&= (DA(\boldsymbol{y})(\boldsymbol{u})B(\boldsymbol{y}+\boldsymbol{u})+A(\boldsymbol{y})DB(\boldsymbol{y})(\boldsymbol{u})+\boldsymbol{o}(\boldsymbol{u}))(\boldsymbol{w}) \\
&= (DA(\boldsymbol{y})(\boldsymbol{u})B(\boldsymbol{y})+A(\boldsymbol{y})DB(\boldsymbol{y})(\boldsymbol{u})+\boldsymbol{o}(\boldsymbol{u}))(\boldsymbol{w})
\end{aligned}$$

Then

$$\boldsymbol{u} \to (DA(\boldsymbol{y})(\boldsymbol{u})B(\boldsymbol{y})+A(\boldsymbol{y})DB(\boldsymbol{y})(\boldsymbol{u}))(\boldsymbol{w})$$

is clearly linear and

$$(\boldsymbol{u},\boldsymbol{w}) \to (DA(\boldsymbol{y})(\boldsymbol{u})B(\boldsymbol{y})+A(\boldsymbol{y})DB(\boldsymbol{y})(\boldsymbol{u}))(\boldsymbol{w})$$

is bilinear and continuous as a function of $\boldsymbol{y}$. By this we mean that for a fixed choice of $(\boldsymbol{u},\boldsymbol{w})$ the resulting Y valued function just described is continuous. Now if each of AB, DA, DB can be differentiated, you could replace $\boldsymbol{y}$ with $\boldsymbol{y}+\hat{\boldsymbol{u}}$ and do a similar computation to obtain as many differentiations as desired, the k^{th} differentiation yielding a k linear function. You can do this as long as A and B have derivatives. Now in the case of the implicit function theorem, you have

$$D\boldsymbol{x}(\boldsymbol{y}) = -D_1\boldsymbol{f}(\boldsymbol{x}(\boldsymbol{y}),\boldsymbol{y})^{-1} D_2\boldsymbol{f}(\boldsymbol{x}(\boldsymbol{y}),\boldsymbol{y}). \tag{8.14}$$

By Lemma 8.1.5 and the implicit function theorem and the chain rule, this is the situation just discussed. Thus $D^2\boldsymbol{x}(\boldsymbol{y})$ can be obtained. Then the formula for it will only involve $D\boldsymbol{x}$ which is known to be continuous. Thus one can continue in this way finding derivatives till $\boldsymbol{f}$ fails to have them. The inverse map never creates difficulties because it is differentiable of order m for any m thanks to Lemma 8.1.5. Thus one can conclude the following corollary.

Corollary 8.2.1 *In the implicit and inverse function theorems, you can replace C^1 with C^k in the statements of the theorems for any $k \in \mathbb{N}$.*

8.3 The Case of $\mathbb{R}^n$

In many applications of the implicit function theorem,

$$\boldsymbol{f} : U \subseteq \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$$

and $\boldsymbol{f}(\boldsymbol{x}_0,\boldsymbol{y}_0) = \boldsymbol{0}$ while $\boldsymbol{f}$ is C^1. How can you recognize the condition of the implicit function theorem which says $D_1\boldsymbol{f}(\boldsymbol{x}_0,\boldsymbol{y}_0)^{-1}$ exists? This is really not hard. You recall the matrix of the transformation $D_1\boldsymbol{f}(\boldsymbol{x}_0,\boldsymbol{y}_0)$ with respect to the usual basis vectors is

$$\begin{pmatrix} f_{1,x_1}(\boldsymbol{x}_0,\boldsymbol{y}_0) & \cdots & f_{1,x_n}(\boldsymbol{x}_0,\boldsymbol{y}_0) \\ \vdots & & \vdots \\ f_{n,x_1}(\boldsymbol{x}_0,\boldsymbol{y}_0) & \cdots & f_{n,x_n}(\boldsymbol{x}_0,\boldsymbol{y}_0) \end{pmatrix}$$

and so $D_1\boldsymbol{f}(\boldsymbol{x}_0,\boldsymbol{y}_0)^{-1}$ exists exactly when the determinant of the above matrix is nonzero. This is the condition to check. In the general case, you just need to verify $D_1\boldsymbol{f}(\boldsymbol{x}_0,\boldsymbol{y}_0)$ is one to one and this can also be accomplished by looking at the matrix of the transformation with respect to some bases on X and Z.

8.4 Exercises

1. Let $A \in \mathscr{L}(X,Y)$. Let $f(x) \equiv Ax$. Verify from the definition that $Df(x) = A$. What if $f(x) = y + Ax$? Note the similarity with functions of a single variable.

2. You have a level surface given by

$$f(x,y,z) = 0,\ f \text{ is } C^1(U), (x,y,z) \in U,$$

The question is whether this deserves to be called a surface. Using the implicit function theorem, show that if $f(x_0,y_0,z_0) = 0$ and if $\frac{\partial f}{\partial z}(x_0,y_0,z_0) \neq 0$ then in some open subset of $\mathbb{R}^3$, the relation $f(x,y,z) = 0$ can be "solved" for z getting say $z = z(x,y)$ such that $f(x,y,z(x,y)) = 0$. What happens if $\frac{\partial f}{\partial x}(x_0,y_0,z_0) \neq 0$ or $\frac{\partial f}{\partial y}(x_0,y_0,z_0) \neq 0$? Explain why z is a C^1 map for (x,y) in some open set.

3. Let $\boldsymbol{x}(t) = (x(t), y(t), z(t))^T$ be a vector valued function defined for $t \in (a,b)$. Then $D\boldsymbol{x}(t) \in \mathscr{L}(\mathbb{R}, \mathbb{R}^3)$. We usually denote this simply as $\boldsymbol{x}'(t)$. Thus, considered as a matrix, it is the 3×1 matrix $(x'(t), y'(t), z'(t))^T$ the T indicating that you take the transpose. Don't worry too much about this. You can also consider this as a vector. What is the geometric significance of this vector? The answer is that this vector is tangent to the curve traced out by $\boldsymbol{x}(t)$ for $t \in (a,b)$. Explain why this is so using the definition of the derivative. You need to describe what is meant by being tangent first. By saying that the line $\boldsymbol{x} = \boldsymbol{a} + t\boldsymbol{b}$ is tangent to a parametric curve consisting of points traced out by $\boldsymbol{x}(t)$ for $t \in (-\delta, \delta)$ at the point $\boldsymbol{a} = \boldsymbol{x}(t)$ which is on both the line and the curve, you would want to have

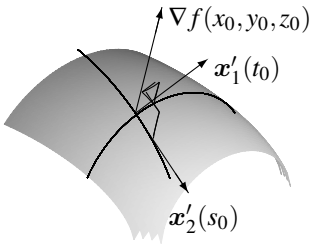
$$\lim_{u \to 0} \frac{\boldsymbol{x}(t+u) - (\boldsymbol{a} + \boldsymbol{b}u)}{u} = \boldsymbol{0}$$

With this definition of what it means for a line to be tangent, explain why the line $\boldsymbol{x}(t) + \boldsymbol{x}'(t)u$ for $u \in (-\delta, \delta)$ is tangent to the curve determined by $t \to \boldsymbol{x}(t)$ at the point $\boldsymbol{x}(t)$. So why would you take the above as a definition of what it means to be tangent? Consider the component functions of $\boldsymbol{x}(t)$. What does the above limit say about the component functions and the corresponding components of $\boldsymbol{b}$ in terms of slopes of lines tangent to curves?

4. Let $f(x,y,z)$ be a C^1 function $f : U \to \mathbb{R}$ where U is an open set in $\mathbb{R}^3$. The gradient vector, defined as

$$\left(\frac{\partial f}{\partial x}(x,y,z) \quad \frac{\partial f}{\partial y}(x,y,z) \quad \frac{\partial f}{\partial z}(x,y,z) \right)^T$$

has fundamental geometric significance illustrated by the following picture.



The way we present this in engineering math is to consider a smooth C^1 curve $(x(t),y(t),z(t))$ for $t \in (a,b)$ such that when $t = c \in (a,b)$, $(x(c),y(c),z(c))$ equals the point (x,y,z) in the level surface and such that $(x(t),y(t),z(t))$ lies in this surface. Then $0 = f(x(t),y(t),z(t))$. Show, using the chain rule, that the gradient vector at the point (x,y,z) is perpendicular to

$$\left(x'(c),y'(c),z'(c)\right).$$

Recall that the chain rule says that for $h(t) = f(x(t),y(t),z(t)), Dh(t) =$

$$\left(\begin{array}{ccc} \frac{\partial f}{\partial x}(x(t),y(t),z(t)) & \frac{\partial f}{\partial y}(x(t),y(t),z(t)) & \frac{\partial f}{\partial z}(x(t),y(t),z(t)) \end{array} \right) \left(\begin{array}{c} x' \\ y' \\ z' \end{array} \right)$$

Since this holds for all such smooth curves in the surface which go through the given point, we say that the gradient vector is perpendicular to the surface. In the picture, there are two intersecting curves which are shown to intersect at a point of the surface. We present this to the students in engineering math and everyone is happy with it, but the argument is specious. Why do there exist any smooth curves in the surface through a point? What would you need to assume to justify the existence of smooth curves in the surface at some point of the level surface? Why?

5. This problem illustrates what can happen when the gradient of a scalar valued function vanishes or is not well defined. Consider the level surface given by

$$z - \sqrt{(x^2+y^2)} = 0.$$

Sketch the graph of this surface. Why is there no unique tangent plane at the origin $(0,0,0)$? Next consider $z^2 - \left(x^2+y^2\right) = 0$. What about a well defined tangent plane at $(0,0,0)$?

6. Suppose you have two level surfaces $f(x,y,z) = 0$ and $g(x,y,z) = 0$ which intersect at a point (x_0,y_0,z_0), each f,g is C^1. Use the implicit function theorem to give conditions which will guarantee that the intersection of these two surfaces near this point is a curve. Explain why.

7. Let X,Y be Banach spaces and let U be an open subset of X. Let $f : U \to Y$ be $C^1(U)$, let $x_0 \in U$, and $\delta > 0$ be given. Show there exists $\varepsilon > 0$ such that if $x_1, x_2 \in B(x_0,\varepsilon)$, then

$$\|f(x_1) - (f(x_2) + Df(x_0)(x_1 - x_2))\| \leq \delta \|x_1 - x_2\|$$

Hint: You know $f(x_1) - f(x_2) = Df(x_2)(x_1 - x_2) + o(x_1 - x_2)$. Use continuity.

8. ↑This problem illustrates how if $Df(x_0)$ is one to one, then near x_0 the same is true of f. Suppose in this problem that all normed linear spaces are finite dimensional. Suppose $Df(x_0)$ is one to one. Here $f : U \to Y$ where $U \subseteq X$.

   (a) Show that there exists $r > 0$ such that $\|Df(x_0)x\| \geq r\|x\|$. To do this, recall equivalence of norms.

   (b) Use the above problem to show that there is $\varepsilon > 0$ such that f is one to one on $B(x_0,\varepsilon)$ provided $Df(x_0)$ is one to one.

9. If U,V are open sets in Banach spaces X,Y respectively and $f:U\to V$ is one to one and onto and both f,f^{-1} are C^1, show that $Df(x):X\to Y$ is one to one and onto for each $x\in U$. **Hint:** $f\circ f^{-1}=identity$. Now use chain rule.

10. A function $f:U\subseteq\mathbb{C}\to\mathbb{C}$ where U is an open set subset of the complex numbers $\mathbb{C}$ is called analytic if

$$\lim_{h\to 0}\frac{f(z+h)-f(z)}{h}\equiv f'(z),\ z=x+iy$$

exists and $z\to f'(z)$ is continuous. Show that if f is analytic on an open set U and if $f'(z)\neq 0$, then there is an open set V containing z such that $f(V)$ is open, f is one to one, and f,f^{-1} are both continuous. **Hint:** This follows very easily from the inverse function theorem. Recall that we have allowed for the field of scalars the complex numbers.

11. Problem 8 has to do with concluding that f is locally one to one if $Df(x_0)$ is only known to be one to one. The next obvious question concerns the situation where $Df(x_0)$ maybe is possibly not one to one but is onto. There are two parts, a linear algebra consideration, followed by an application of the inverse function theorem. Thus these two problems are each generalizations of the inverse function theorem.

    (a) Suppose X is a finite dimensional vector space and $M\in\mathscr{L}(X,Y)$ is onto Y. Consider a basis for $M(X)=Y,\{Mx_1,\cdots,Mx_n\}$. Verify that $\{x_1,\cdots,x_n\}$ is linearly independent. Define $\hat{X}\equiv\operatorname{span}(x_1,\cdots,x_n)$. Show that if $\hat{M}$ is the restriction of M to $\hat{X}$, then M is one to one and onto Y.

    (b) Now suppose $f:U\subseteq X\to Y$ is C^1 and $Df(x_0)$ is onto Y. Show that there is a ball $B(f(x_0),\varepsilon)$ and an open set $V\subseteq X$ such that $f(V)\supseteq B(f(x_0),\varepsilon)$ so that if $Df(x)$ is onto for each $x\in U$, then $f(U)$ is an open set. This is called the open map theorem. You might use the inverse function theorem with the spaces $\hat{X},Y$. You might want to consider Problem 1. This is a nice illustration of why we developed the inverse and implicit function theorems on arbitrary normed linear spaces. You will see that this is a fairly easy problem.

12. Recall that a function $f:U\subseteq X\to Y$ where here assume X is finite dimensional, is Gateaux differentiable if

$$\lim_{t\to 0}\frac{f(x+tv)-f(x)}{t}\equiv D_vf(x)$$

exists. Here $t\in\mathbb{R}$. Suppose that $x\to D_vf(x)$ exists and is continuous on U. Show it follows that f is differentiable and in fact $D_vf(x)=Df(x)v$. **Hint:** Let $g(\boldsymbol{y})\equiv f(\sum_i y_ix_i)$ and argue that the partial derivatives of g all exist and are continuous. Conclude that g is C^1 and then argue that f is just the composition of g with a linear map.

13. Let

$$f(x,y)=\begin{cases}\frac{(x^2-y^4)^2}{(x^2+y^4)^2}\text{ if }(x,y)\neq(0,0)\\ 1\text{ if }(x,y)=(0,0)\end{cases}$$

Show that f is not differentiable, and in fact is not even continuous, but $D_{\boldsymbol{v}}f(0,0)$ exists and equals 0 for every $\boldsymbol{v}\neq\mathbf{0}$.

14. Let
$$f(x,y) = \begin{cases} \frac{xy^4}{x^2+y^8} \text{ if } (x,y) \neq (0,0) \\ 0 \text{ if } (x,y) = (0,0) \end{cases}$$
Show that f is not differentiable, and in fact is not even continuous, but $D_{\boldsymbol{v}} f(0,0)$ exists and equals 0 for every $\boldsymbol{v} \neq \boldsymbol{0}$.

8.5 The Method of Lagrange Multipliers

As an application of the implicit function theorem, consider the method of Lagrange multipliers from calculus. Recall the problem is to maximize or minimize a function subject to equality constraints. Let $f : U \to \mathbb{R}$ be a C^1 function where $U \subseteq \mathbb{R}^n$ and let
$$g_i(\boldsymbol{x}) = 0,\ i = 1,\cdots,m \tag{8.15}$$
be a collection of equality constraints with $m < n$. Now consider the system of nonlinear equations
$$\begin{aligned} f(\boldsymbol{x}) &= a \\ g_i(\boldsymbol{x}) &= 0,\ i = 1,\cdots,m. \end{aligned}$$
$\boldsymbol{x}_0$ is a local maximum if $f(\boldsymbol{x}_0) \geq f(\boldsymbol{x})$ for all $\boldsymbol{x}$ near $\boldsymbol{x}_0$ which also satisfies the constraints 8.15. A local minimum is defined similarly. Let $\boldsymbol{F} : U \times \mathbb{R} \to \mathbb{R}^{m+1}$ be defined by
$$\boldsymbol{F}(\boldsymbol{x},a) \equiv \begin{pmatrix} f(\boldsymbol{x}) - a \\ g_1(\boldsymbol{x}) \\ \vdots \\ g_m(\boldsymbol{x}) \end{pmatrix}. \tag{8.16}$$
Now consider the $m+1 \times n$ Jacobian matrix, the matrix of the linear transformation, $D_1 \boldsymbol{F}(\boldsymbol{x},a)$ with respect to the usual basis for $\mathbb{R}^n$ and $\mathbb{R}^{m+1}$.
$$\begin{pmatrix} f_{x_1}(\boldsymbol{x}_0) & \cdots & f_{x_n}(\boldsymbol{x}_0) \\ g_{1x_1}(\boldsymbol{x}_0) & \cdots & g_{1x_n}(\boldsymbol{x}_0) \\ \vdots & & \vdots \\ g_{mx_1}(\boldsymbol{x}_0) & \cdots & g_{mx_n}(\boldsymbol{x}_0) \end{pmatrix}.$$
If this matrix has rank $m+1$ then some $m+1 \times m+1$ submatrix has nonzero determinant. It follows from the implicit function theorem that there exist $m+1$ variables, $x_{i_1},\cdots,x_{i_{m+1}}$ such that the system
$$\boldsymbol{F}(\boldsymbol{x},a) = \boldsymbol{0} \tag{8.17}$$
specifies these $m+1$ variables as a function of the remaining $n-(m+1)$ variables and a in an open set of $\mathbb{R}^{n-m}$. Thus there is a solution $(\boldsymbol{x},a)$ to 8.17 for some $\boldsymbol{x}$ close to $\boldsymbol{x}_0$ whenever a is in some open interval. Therefore, $\boldsymbol{x}_0$ cannot be either a local minimum or a local maximum. It follows that if $\boldsymbol{x}_0$ is either a local maximum or a local minimum, then the above matrix must have rank less than $m+1$ which requires the rows to be linearly dependent. Thus, there exist m scalars,
$$\lambda_1,\cdots,\lambda_m,$$

and a scalar μ, not all zero such that

$$\mu \begin{pmatrix} f_{x_1}(\boldsymbol{x}_0) \\ \vdots \\ f_{x_n}(\boldsymbol{x}_0) \end{pmatrix} = \lambda_1 \begin{pmatrix} g_{1x_1}(\boldsymbol{x}_0) \\ \vdots \\ g_{1x_n}(\boldsymbol{x}_0) \end{pmatrix} + \cdots + \lambda_m \begin{pmatrix} g_{mx_1}(\boldsymbol{x}_0) \\ \vdots \\ g_{mx_n}(\boldsymbol{x}_0) \end{pmatrix}. \tag{8.18}$$

If the column vectors

$$\begin{pmatrix} g_{1x_1}(\boldsymbol{x}_0) \\ \vdots \\ g_{1x_n}(\boldsymbol{x}_0) \end{pmatrix}, \cdots \begin{pmatrix} g_{mx_1}(\boldsymbol{x}_0) \\ \vdots \\ g_{mx_n}(\boldsymbol{x}_0) \end{pmatrix} \tag{8.19}$$

are linearly independent, then, $\mu \neq 0$ and dividing by μ yields an expression of the form

$$\begin{pmatrix} f_{x_1}(\boldsymbol{x}_0) \\ \vdots \\ f_{x_n}(\boldsymbol{x}_0) \end{pmatrix} = \lambda_1 \begin{pmatrix} g_{1x_1}(\boldsymbol{x}_0) \\ \vdots \\ g_{1x_n}(\boldsymbol{x}_0) \end{pmatrix} + \cdots + \lambda_m \begin{pmatrix} g_{mx_1}(\boldsymbol{x}_0) \\ \vdots \\ g_{mx_n}(\boldsymbol{x}_0) \end{pmatrix} \tag{8.20}$$

at every point $\boldsymbol{x}_0$ which is either a local maximum or a local minimum. This proves the following theorem.

Theorem 8.5.1 *Let U be an open subset of $\mathbb{R}^n$ and let $f : U \to \mathbb{R}$ be a C^1 function. Then if $\boldsymbol{x}_0 \in U$ is either a local maximum or local minimum of f subject to the constraints 8.15, then 8.18 must hold for some scalars $\mu, \lambda_1, \cdots, \lambda_m$ not all equal to zero. If the vectors in 8.19 are linearly independent, it follows that an equation of the form 8.20 holds.*

8.6 The Taylor Formula

First recall the Taylor formula with the Lagrange form of the remainder. It will only be needed on $[0,1]$ so that is what I will show.

Theorem 8.6.1 *Let $h : [0,1] \to \mathbb{R}$ have $m+1$ derivatives. Then there exists $t \in (0,1)$ such that*

$$h(1) = h(0) + \sum_{k=1}^{m} \frac{h^{(k)}(0)}{k!} + \frac{h^{(m+1)}(t)}{(m+1)!}.$$

Proof: Let K be a number chosen such that

$$h(1) - \left(h(0) + \sum_{k=1}^{m} \frac{h^{(k)}(0)}{k!} + K \right) = 0$$

Now the idea is to find K. To do this, let

$$F(t) = h(1) - \left(h(t) + \sum_{k=1}^{m} \frac{h^{(k)}(t)}{k!} (1-t)^k + K(1-t)^{m+1} \right)$$

Then $F(1) = F(0) = 0$. Therefore, by Rolle's theorem or the mean value theorem, Theorem 5.8.3, there exists t between 0 and 1 such that $F'(t) = 0$. Thus,

$$\begin{aligned} 0 &= -F'(t) = h'(t) + \sum_{k=1}^{m} \frac{h^{(k+1)}(t)}{k!} (1-t)^k \\ &\quad - \sum_{k=1}^{m} \frac{h^{(k)}(t)}{k!} k(1-t)^{k-1} - K(m+1)(1-t)^m \end{aligned}$$

And so

$$\begin{aligned} &= h'(t)+\sum_{k=1}^{m}\frac{h^{(k+1)}(t)}{k!}(1-t)^k-\sum_{k=0}^{m-1}\frac{h^{(k+1)}(t)}{k!}(1-t)^k \\ &\quad -K(m+1)(1-t)^m \\ &= h'(t)+\frac{h^{(m+1)}(t)}{m!}(1-t)^m-h'(t)-K(m+1)(1-t)^m \end{aligned}$$

and so $K=\frac{h^{(m+1)}(t)}{(m+1)!}$.This proves the theorem. ■

Now let $f:U\to\mathbb{R}$ where $U\subseteq X$ a normed vector space and suppose $f\in C^m(U)$ and suppose $D^{m+1}f(\boldsymbol{x})$ exists on U. Let $\boldsymbol{x}\in U$ and let $r>0$ be such that

$$B(\boldsymbol{x},r)\subseteq U.$$

Then for $\|\boldsymbol{v}\|<r$ consider

$$f(\boldsymbol{x}+t\boldsymbol{v})-f(\boldsymbol{x})\equiv h(t)$$

for $t\in[0,1]$. Then by the chain rule,

$$h'(t)=Df(\boldsymbol{x}+t\boldsymbol{v})(\boldsymbol{v});\ h''(t)=D^2f(\boldsymbol{x}+t\boldsymbol{v})(\boldsymbol{v})(\boldsymbol{v})$$

and continuing in this way,

$$h^{(k)}(t)=D^{(k)}f(\boldsymbol{x}+t\boldsymbol{v})(\boldsymbol{v})(\boldsymbol{v})\cdots(\boldsymbol{v})\equiv D^{(k)}f(\boldsymbol{x}+t\boldsymbol{v})\boldsymbol{v}^k.$$

It follows from Taylor's formula for a function of one variable given above that

$$f(\boldsymbol{x}+\boldsymbol{v})=f(\boldsymbol{x})+\sum_{k=1}^{m}\frac{D^{(k)}f(\boldsymbol{x})\boldsymbol{v}^k}{k!}+\frac{D^{(m+1)}f(\boldsymbol{x}+t\boldsymbol{v})\boldsymbol{v}^{m+1}}{(m+1)!}. \tag{8.21}$$

This proves the following theorem.

Theorem 8.6.2 *Let $f:U\to\mathbb{R}$ and let $f\in C^{m+1}(U)$. Then if*

$$B(\boldsymbol{x},r)\subseteq U,$$

and $\|\boldsymbol{v}\|<r$, there exists $t\in(0,1)$ such that 8.21 holds.

8.7 Second Derivative Test

Now consider the case where $U\subseteq\mathbb{R}^n$ and $f:U\to\mathbb{R}$ is $C^2(U)$. Then from Taylor's theorem, if $\boldsymbol{v}$ is small enough, there exists $t\in(0,1)$ such that

$$f(\boldsymbol{x}+\boldsymbol{v})=f(\boldsymbol{x})+Df(\boldsymbol{x})\boldsymbol{v}+\frac{D^2f(\boldsymbol{x}+t\boldsymbol{v})\boldsymbol{v}^2}{2}. \tag{8.22}$$

Consider

$$\begin{aligned} D^2f(\boldsymbol{x}+t\boldsymbol{v})(\boldsymbol{e}_i)(\boldsymbol{e}_j) &\equiv D(D(f(\boldsymbol{x}+t\boldsymbol{v}))\boldsymbol{e}_i)\boldsymbol{e}_j \\ &= D\left(\frac{\partial f(\boldsymbol{x}+t\boldsymbol{v})}{\partial x_i}\right)\boldsymbol{e}_j=\frac{\partial^2 f(\boldsymbol{x}+t\boldsymbol{v})}{\partial x_j\partial x_i} \end{aligned}$$

where $\boldsymbol{e}_i$ are the usual basis vectors. Lettin $\boldsymbol{v} = \sum_{i=1}^n v_i \boldsymbol{e}_i$, the second derivative term in 8.22 reduces to

$$\frac{1}{2}\sum_{i,j} D^2 f(\boldsymbol{x}+t\boldsymbol{v})(\boldsymbol{e}_i)(\boldsymbol{e}_j) v_i v_j = \frac{1}{2}\sum_{i,j} H_{ij}(\boldsymbol{x}+t\boldsymbol{v}) v_i v_j$$

where

$$H_{ij}(\boldsymbol{x}+t\boldsymbol{v}) = D^2 f(\boldsymbol{x}+t\boldsymbol{v})(\boldsymbol{e}_i)(\boldsymbol{e}_j) = \frac{\partial^2 f(\boldsymbol{x}+t\boldsymbol{v})}{\partial x_j \partial x_i}.$$

Definition 8.7.1 *The matrix whose ij^{th} entry is $\frac{\partial^2 f(\boldsymbol{x})}{\partial x_j \partial x_i}$ is called the Hessian matrix, denoted as $\boldsymbol{H}(\boldsymbol{x})$.*

From Theorem 7.10.1, this is a symmetric real matrix, thus self adjoint. By the continuity of the second partial derivative,

$$f(\boldsymbol{x}+\boldsymbol{v}) = f(\boldsymbol{x}) + Df(\boldsymbol{x})\boldsymbol{v} + \frac{1}{2}\boldsymbol{v}^T H(\boldsymbol{x})\boldsymbol{v} +$$

$$\frac{1}{2}\left(\boldsymbol{v}^T\left(H(\boldsymbol{x}+t\boldsymbol{v}) - H(\boldsymbol{x})\right)\boldsymbol{v}\right). \tag{8.23}$$

where the last two terms involve ordinary matrix multiplication and

$$\boldsymbol{v}^T = \begin{pmatrix} v_1 & \cdots & v_n \end{pmatrix}$$

for v_i the components of $\boldsymbol{v}$ relative to the standard basis.

Definition 8.7.2 *Let $f : D \to \mathbb{R}$ where D is a subset of some normed vector space. Then f has a local minimum at $\boldsymbol{x} \in D$ if there exists $\delta > 0$ such that for all $\boldsymbol{y} \in B(\boldsymbol{x}, \delta)$*

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}).$$

f has a local maximum at $\boldsymbol{x} \in D$ if there exists $\delta > 0$ such that for all $\boldsymbol{y} \in B(\boldsymbol{x}, \delta)$

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}).$$

Theorem 8.7.3 *If $f : U \to \mathbb{R}$ where U is an open subset of $\mathbb{R}^n$ and f is C^2, suppose $Df(\boldsymbol{x}) = 0$. Then if $H(\boldsymbol{x})$ has all positive eigenvalues, $\boldsymbol{x}$ is a local minimum. If the Hessian matrix $H(\boldsymbol{x})$ has all negative eigenvalues, then $\boldsymbol{x}$ is a local maximum. If $H(\boldsymbol{x})$ has a positive eigenvalue, then there exists a direction in which f has a local minimum at $\boldsymbol{x}$, while if $H(\boldsymbol{x})$ has a negative eigenvalue, there exists a direction in which $H(\boldsymbol{x})$ has a local maximum at $\boldsymbol{x}$.*

Proof: Since $Df(\boldsymbol{x}) = 0$, formula 8.23 holds and by continuity of the second derivative, $H(\boldsymbol{x})$ is a symmetric matrix. Thus $H(\boldsymbol{x})$ has all real eigenvalues. Suppose first that $H(\boldsymbol{x})$ has all positive eigenvalues and that all are larger than $\delta^2 > 0$. Then by Theorem 1.4.1, $H(\boldsymbol{x})$ has an orthonormal basis of eigenvectors, $\{\boldsymbol{v}_i\}_{i=1}^n$ and if $\boldsymbol{u}$ is an arbitrary vector, such that $\boldsymbol{u} = \sum_{j=1}^n u_j \boldsymbol{v}_j$ where $u_j = \boldsymbol{u} \cdot \boldsymbol{v}_j$, then

$$\boldsymbol{u}^T H(\boldsymbol{x})\,\boldsymbol{u} = \sum_{j=1}^n u_j \boldsymbol{v}_j^T H(\boldsymbol{x}) \sum_{j=1}^n u_j \boldsymbol{v}_j$$

$$= \sum_{j=1}^{n} u_j^2 \lambda_j \geq \delta^2 \sum_{j=1}^{n} u_j^2 = \delta^2 |\boldsymbol{u}|^2 .$$

From 8.23 and the continuity of H, if $\boldsymbol{v}$ is small enough,

$$f(\boldsymbol{x}+\boldsymbol{v}) \geq f(\boldsymbol{x}) + \frac{1}{2}\delta^2 |\boldsymbol{v}|^2 - \frac{1}{4}\delta^2 |\boldsymbol{v}|^2 = f(\boldsymbol{x}) + \frac{\delta^2}{4} |\boldsymbol{v}|^2 .$$

This shows the first claim of the theorem. The second claim follows from similar reasoning. Suppose $H(\boldsymbol{x})$ has a positive eigenvalue λ^2. Then let $\boldsymbol{v}$ be an eigenvector for this eigenvalue. Then from 8.23,

$$f(\boldsymbol{x}+t\boldsymbol{v}) = f(\boldsymbol{x}) + \frac{1}{2}t^2 \boldsymbol{v}^T H(\boldsymbol{x})\boldsymbol{v} +$$

$$\frac{1}{2}t^2 \left(\boldsymbol{v}^T \left(H(\boldsymbol{x}+t\boldsymbol{v}) - H(\boldsymbol{x})\right)\boldsymbol{v}\right)$$

which implies

$$\begin{aligned} f(\boldsymbol{x}+t\boldsymbol{v}) &= f(\boldsymbol{x}) + \frac{1}{2}t^2\lambda^2 |\boldsymbol{v}|^2 + \frac{1}{2}t^2 \left(\boldsymbol{v}^T \left(H(\boldsymbol{x}+t\boldsymbol{v}) - H(\boldsymbol{x})\right)\boldsymbol{v}\right) \\ &\geq f(\boldsymbol{x}) + \frac{1}{4}t^2\lambda^2 |\boldsymbol{v}|^2 \end{aligned}$$

whenever t is small enough. Thus in the direction $\boldsymbol{v}$ the function has a local minimum at $\boldsymbol{x}$. The assertion about the local maximum in some direction follows similarly. This proves the theorem. ■

This theorem is an analogue of the second derivative test for higher dimensions. As in one dimension, when there is a zero eigenvalue, it may be impossible to determine from the Hessian matrix what the local qualitative behavior of the function is. For example, consider

$$f_1(x,y) = x^4 + y^2, \; f_2(x,y) = -x^4 + y^2.$$

Then $Df_i(0,0) = \mathbf{0}$ and for both functions, the Hessian matrix evaluated at $(0,0)$ equals

$$\begin{pmatrix} 0 & 0 \\ 0 & 2 \end{pmatrix}$$

but the behavior of the two functions is very different near the origin. The second has a saddle point while the first has a minimum there.

8.8 The Rank Theorem

This is a very interesting result. The proof follows Marsden and Hoffman. First here is some linear algebra.

Theorem 8.8.1 *Let $L : \mathbb{R}^n \to \mathbb{R}^N$ have rank m. Then there exists a basis*

$$\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m, \boldsymbol{u}_{m+1}, \cdots, \boldsymbol{u}_n\}$$

such that a basis for $\ker(L)$ *is* $\{\boldsymbol{u}_{m+1}, \cdots, \boldsymbol{u}_n\}$.

Proof: Since L has rank m, there is a basis for $L(\mathbb{R}^n)$ which is of the form

$$\{L\boldsymbol{u}_1, \cdots, L\boldsymbol{u}_m\}$$

Then if $\sum_i c_i \boldsymbol{u}_i = 0$ you can do L to both sides and conclude that each $c_i = 0$. Hence $\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m\}$ is linearly independent. Let $\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_k\}$ be a basis for $\ker(L)$. Let $\boldsymbol{x} \in \mathbb{R}^n$. Then $L\boldsymbol{x} = \sum_{i=1}^m c_i L\boldsymbol{u}_i$ for some choice of scalars c_i. Hence $L(\boldsymbol{x} - \sum_{i=1}^m c_i \boldsymbol{u}_i) = \boldsymbol{0}$ which shows that there exist d_j such that $\boldsymbol{x} = \sum_{i=1}^m c_i \boldsymbol{u}_i + \sum_{j=1}^k d_j \boldsymbol{v}_j$ It follows that

$$\operatorname{span}(\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m, \boldsymbol{v}_1, \cdots, \boldsymbol{v}_k) = \mathbb{R}^n$$

Is this set of vectors linearly independent? Suppose $\sum_{i=1}^m c_i \boldsymbol{u}_i + \sum_{j=1}^k d_j \boldsymbol{v}_j = \boldsymbol{0}$ Do L to both sides to get $\sum_{i=1}^m c_i L\boldsymbol{u}_i = \boldsymbol{0}$ Thus each $c_i = 0$. Hence $\sum_{j=1}^k d_j \boldsymbol{v}_j = \boldsymbol{0}$ and so each $d_j = 0$ also. It follows that $k = n - m$ and we can let

$$\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_k\} = \{\boldsymbol{u}_{m+1}, \cdots, \boldsymbol{u}_n\}. \blacksquare$$

Another useful linear algebra result is the following lemma.

Lemma 8.8.2 *Let $V \subseteq \mathbb{R}^n$ be a subspace and suppose $A(\boldsymbol{x}) \in \mathscr{L}(V, \mathbb{R}^N)$ for $\boldsymbol{x}$ in some open set U. Also suppose $\boldsymbol{x} \to A(\boldsymbol{x})$ is continuous for $\boldsymbol{x} \in U$. Then if $A(\boldsymbol{x}_0)$ is one to one on V for some $\boldsymbol{x}_0 \in U$, then it follows that for all $\boldsymbol{x}$ close enough to $\boldsymbol{x}_0$, $A(\boldsymbol{x})$ is also one to one on V.*

Proof: Consider V as an inner product space with the inner product from $\mathbb{R}^n$ and $A(\boldsymbol{x})^* A(\boldsymbol{x})$. Then $A(\boldsymbol{x})^* A(\boldsymbol{x}) \in \mathscr{L}(V, V)$ and $\boldsymbol{x} \to A(\boldsymbol{x})^* A(\boldsymbol{x})$ is also continuous. Also for $\boldsymbol{v} \in V$,

$$\left(A(\boldsymbol{x})^* A(\boldsymbol{x})\, \boldsymbol{v}, \boldsymbol{v}\right)_V = \left(A(\boldsymbol{x})\, \boldsymbol{v}, A(\boldsymbol{x})\, \boldsymbol{v}\right)_{\mathbb{R}^N}$$

If $A(\boldsymbol{x}_0)^* A(\boldsymbol{x}_0)\, \boldsymbol{v} = \boldsymbol{0}$, then from the above, it follows that $A(\boldsymbol{x}_0)\, \boldsymbol{v} = \boldsymbol{0}$ also. Therefore, $\boldsymbol{v} = \boldsymbol{0}$ and so $A(\boldsymbol{x}_0)^* A(\boldsymbol{x}_0)$ is one to one on V. For all $\boldsymbol{x}$ close enough to $\boldsymbol{x}_0$, it follows from continuity that $A(\boldsymbol{x})^* A(\boldsymbol{x})$ is also one to one. Thus, for such $\boldsymbol{x}$, if $A(\boldsymbol{x})\, \boldsymbol{v} = \boldsymbol{0}$, Then $A(\boldsymbol{x})^* A(\boldsymbol{x})\, \boldsymbol{v} = \boldsymbol{0}$ and so $\boldsymbol{v} = \boldsymbol{0}$. Thus, for $\boldsymbol{x}$ close enough to $\boldsymbol{x}_0$, it follows that $A(\boldsymbol{x})$ is also one to one on V. $\blacksquare$

Theorem 8.8.3 *Let $\boldsymbol{f} : A \subseteq \mathbb{R}^n \to \mathbb{R}^N$ where A is open in $\mathbb{R}^n$. Let $\boldsymbol{f}$ be a C^r function and suppose that $D\boldsymbol{f}(\boldsymbol{x})$ has rank m for all $\boldsymbol{x} \in A$. Let $\boldsymbol{x}_0 \in A$. Then there are open sets $U, V \subseteq \mathbb{R}^n$ with $\boldsymbol{x}_0 \in V$, and a C^r function $\boldsymbol{h} : U \to V$ with inverse $\boldsymbol{h}^{-1} : V \to U$ also C^r such that $\boldsymbol{f} \circ \boldsymbol{h}$ depends only on $(x_1, \cdots, x_m)$.*

Proof: Let $L = D\boldsymbol{f}(x_0)$, and $N_0 = \ker L$. Using the above linear algebra theorem, there exists

$$\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m\}$$

such that $\{L\boldsymbol{u}_1, \cdots, L\boldsymbol{u}_m\}$ is a basis for $L\mathbb{R}^n$. Extend to form a basis for $\mathbb{R}^n$,

$$\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m, \boldsymbol{u}_{m+1}, \cdots, \boldsymbol{u}_n\}$$

such that a basis for $N_0 = \ker L$ is $\{\boldsymbol{u}_{m+1}, \cdots, \boldsymbol{u}_n\}$. Let

$$M \equiv \operatorname{span}(\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m).$$

Let the coordinate maps be ψ_k so that if $\boldsymbol{x} \in \mathbb{R}^n$,

$$\boldsymbol{x} = \psi_1(\boldsymbol{x})\boldsymbol{u}_1 + \cdots + \psi_n(\boldsymbol{x})\boldsymbol{u}_n$$

Since these coordinate maps are linear, they are infinitely differentiable.

Next I will define coordinate maps for $\boldsymbol{x} \in \mathbb{R}^N$. Then by the above construction, $\{L\boldsymbol{u}_1, \cdots, L\boldsymbol{u}_m\}$ is a basis for $L(\mathbb{R}^n)$. Let a basis for $\mathbb{R}^N$ be

$$\{L\boldsymbol{u}_1, \cdots, L\boldsymbol{u}_m, \boldsymbol{v}_{m+1}, \cdots, \boldsymbol{v}_N\}$$

(Note that, since the rank of $D\boldsymbol{f}(\boldsymbol{x}) = m$ you must have $N \geq m$.) The coordinate maps ϕ_i will be defined as follows for $\boldsymbol{x} \in \mathbb{R}^N$.

$$\boldsymbol{x} = \phi_1(\boldsymbol{x})L\boldsymbol{u}_1 + \cdots \phi_m(\boldsymbol{x})L\boldsymbol{u}_m + \phi_{m+1}(\boldsymbol{x})\boldsymbol{v}_{m+1} + \cdots + \phi_N(\boldsymbol{x})\boldsymbol{v}_N$$

Now define two infinitely differentiable maps $G : \mathbb{R}^n \to \mathbb{R}^n$ and $H : \mathbb{R}^N \to \mathbb{R}^n$,

$$G(\boldsymbol{x}) \equiv \left(0, \cdots, 0, \psi_{m+1}(\boldsymbol{x}), \cdots, \psi_n(\boldsymbol{x})\right)$$

$$H(\boldsymbol{y}) \equiv (\phi_1(\boldsymbol{y}), \cdots, \phi_m(\boldsymbol{y}), 0, \cdots, 0)$$

For $\boldsymbol{x} \in A \subseteq \mathbb{R}^n$, let

$$\boldsymbol{g}(\boldsymbol{x}) \equiv H(\boldsymbol{f}(\boldsymbol{x})) + G(\boldsymbol{x}) \in \mathbb{R}^n$$

Thus the first term picks out the first m entries of $\boldsymbol{f}(\boldsymbol{x})$ and the second term the last $n-m$ entries of $\boldsymbol{x}$. It is of the form

$$\left(\phi_1(\boldsymbol{f}(\boldsymbol{x})), \cdots, \phi_m(\boldsymbol{f}(\boldsymbol{x})), \psi_{m+1}(\boldsymbol{x}), \cdots, \psi_n(\boldsymbol{x})\right)$$

Then

$$D\boldsymbol{g}(\boldsymbol{x}_0)(\boldsymbol{v}) = HL(\boldsymbol{v}) + G\,\boldsymbol{v} = HL\boldsymbol{v} + G\boldsymbol{v} \tag{8.24}$$

which is of the form

$$D\boldsymbol{g}(\boldsymbol{x}_0)(\boldsymbol{v}) = \left(\phi_1(L\boldsymbol{v}), \cdots, \phi_m(L\boldsymbol{v}), \psi_{m+1}(\boldsymbol{v}), \cdots, \psi_n(\boldsymbol{v})\right)$$

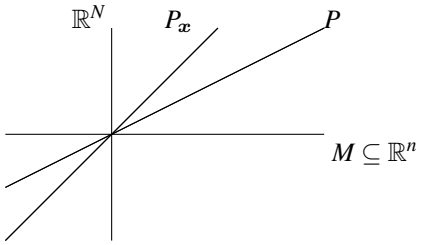
If this equals $\mathbf{0}$, then all the components of $\boldsymbol{v}$, $\psi_{m+1}(\boldsymbol{v}), \cdots, \psi_n(\boldsymbol{v})$ are equal to 0. Hence

$$\boldsymbol{v} = \sum_{i=1}^{m} c_i \boldsymbol{u}_i.$$

But also the coordinates of $L\boldsymbol{v}$, $\phi_1(L\boldsymbol{v}), \cdots, \phi_m(L\boldsymbol{v})$ are all zero so $L\boldsymbol{v} = \mathbf{0}$ and so $\mathbf{0} = \sum_{i=1}^{m} c_i L\boldsymbol{u}_i$ so by independence of the $L\boldsymbol{u}_i$, each $c_i = 0$ and consequently $\boldsymbol{v} = \mathbf{0}$.

This proves the conditions for the inverse function theorem are valid for $\boldsymbol{g}$. Therefore, there is an open ball U and an open set V, $\boldsymbol{x}_0 \in V$, such that $\boldsymbol{g} : V \to U$ is a C^r map and its inverse $\boldsymbol{g}^{-1} : U \to V$ is also. We can assume by continuity and Lemma 8.8.2 that V and U are small enough that for each $\boldsymbol{x} \in V, D\boldsymbol{g}(\boldsymbol{x})$ is one to one. This follows from the fact that $\boldsymbol{x} \to D\boldsymbol{g}(\boldsymbol{x})$ is continuous.

Since it is assumed that $D\boldsymbol{f}(\boldsymbol{x})$ is of rank $m, D\boldsymbol{f}(\boldsymbol{x})(\mathbb{R}^n)$ is a subspace which is m dimensional, denoted as $P_{\boldsymbol{x}}$. Also denote $L(\mathbb{R}^n) = L(M)$ as P.



Thus $\{L\boldsymbol{u}_1, \cdots, L\boldsymbol{u}_m\}$ is a basis for P. Using Lemma 8.8.2 again, by making V, U smaller if necessary, one can also assume that for each $\boldsymbol{x} \in V$, $D\boldsymbol{f}(\boldsymbol{x})$ is one to one on M (although not on $\mathbb{R}^n$) and $HD\boldsymbol{f}(\boldsymbol{x})$ is one to one on M. This follows from continuity and the fact that $L = D\boldsymbol{f}(\boldsymbol{x}_0)$ is one to one on M. Therefore, it is also the case that $D\boldsymbol{f}(\boldsymbol{x})$ maps the m dimensional space M **onto** the m dimensional space $P_{\boldsymbol{x}}$ and H is one to one on $P_{\boldsymbol{x}}$. The reason for this last claim is as follows: If $H\boldsymbol{z} = \boldsymbol{0}$ where $\boldsymbol{z} \in P_{\boldsymbol{x}}$, then $HD\boldsymbol{f}(\boldsymbol{x})\boldsymbol{w} = \boldsymbol{0}$ where $\boldsymbol{w} \in M$ and $D\boldsymbol{f}(\boldsymbol{x})\boldsymbol{w} = \boldsymbol{z}$. Hence $\boldsymbol{w} = \boldsymbol{0}$ because $HD\boldsymbol{f}(\boldsymbol{x})$ is one to one, and so $\boldsymbol{z} = \boldsymbol{0}$ which shows that indeed H is one to one on $P_{\boldsymbol{x}}$.

Denote as $L_{\boldsymbol{x}}$ the inverse of H which is defined on $\mathbb{R}^m \times \boldsymbol{0}$, $L_{\boldsymbol{x}} : \mathbb{R}^m \times \boldsymbol{0} \to P_{\boldsymbol{x}}$. That $\boldsymbol{0}$ refers to the $N-m$ string of zeros in the definition given above for H.

Define $\boldsymbol{h} \equiv \boldsymbol{g}^{-1}$ and consider $\boldsymbol{f}_1 \equiv \boldsymbol{f} \circ \boldsymbol{h}$. It is desired to show that $\boldsymbol{f}_1$ depends only on $x_1, \cdots, x_m$. Let D_1 refer to $(x_1, \cdots, x_m)$ and let D_2 refer to $(x_{m+1}, \cdots, x_n)$. Then $\boldsymbol{f} = \boldsymbol{f}_1 \circ \boldsymbol{g}$ and so by the chain rule

$$D\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{y}) = D\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x}))D\boldsymbol{g}(\boldsymbol{x})(\boldsymbol{y}) \tag{8.25}$$

Now as in 8.24, for $\boldsymbol{y} \in \mathbb{R}^n$,

$$D\boldsymbol{g}(\boldsymbol{x})(\boldsymbol{y}) = HD\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{y}) + G\boldsymbol{y}$$

$$= \left(\phi_1(D\boldsymbol{f}(\boldsymbol{x})\boldsymbol{y}), \cdots, \phi_m(D\boldsymbol{f}(\boldsymbol{x})\boldsymbol{y}), \psi_{m+1}(\boldsymbol{y}), \cdots, \psi_n(\boldsymbol{y})\right)$$

Recall that from the above definitions of H and G,

$$G(\boldsymbol{y}) \equiv \left(0, \cdots, 0, \psi_{m+1}(\boldsymbol{y}), \cdots, \psi_n(\boldsymbol{y})\right)$$

$$H(D\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{y})) = (\phi_1(D\boldsymbol{f}(\boldsymbol{x})\boldsymbol{y}), \cdots, \phi_m(D\boldsymbol{f}(\boldsymbol{x})\boldsymbol{y}), 0, \cdots, 0)$$

Let $\pi_1 : \mathbb{R}^n \to \mathbb{R}^m$ denote the projection onto the first m positions and π_2 the projection onto the last $n-m$. Thus

$$\begin{aligned} \pi_1 D\boldsymbol{g}(\boldsymbol{x})(\boldsymbol{y}) &= (\phi_1(D\boldsymbol{f}(\boldsymbol{x})\boldsymbol{y}), \cdots, \phi_m(D\boldsymbol{f}(\boldsymbol{x})\boldsymbol{y})) \\ \pi_2 D\boldsymbol{g}(\boldsymbol{x})(\boldsymbol{y}) &= \left(\psi_{m+1}(\boldsymbol{y}), \cdots, \psi_n(\boldsymbol{y})\right) \end{aligned}$$

Now in general, for $\boldsymbol{z} \in \mathbb{R}^n$,

$$D\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x}))\boldsymbol{z} = D_1\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x}))\pi_1\boldsymbol{z} + D_2\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x}))\pi_2\boldsymbol{z}$$

Therefore, it follows that $D\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x}))D\boldsymbol{g}(\boldsymbol{x})(\boldsymbol{y})$ is given by

$$\begin{aligned} D\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{y}) &= D\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x}))D\boldsymbol{g}(\boldsymbol{x})(\boldsymbol{y}) \\ &= D_1\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x}))\pi_1 D\boldsymbol{g}(\boldsymbol{x})(\boldsymbol{y}) + D_2\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x}))\pi_2 D\boldsymbol{g}(\boldsymbol{x})(\boldsymbol{y}) \end{aligned}$$

$$\begin{aligned} D\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{y}) &= D\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x}))D\boldsymbol{g}(\boldsymbol{x})(\boldsymbol{y}) = D_1\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x}))\overbrace{\pi_1 HD\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{y})}^{=\pi_1 D\boldsymbol{g}(\boldsymbol{x})(\boldsymbol{y})} \\ &\quad +D_2\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x}))\pi_2 G\boldsymbol{y} \end{aligned}$$

We need to verify the last term equals 0. Solving for this term,

$$D_2\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x}))\pi_2 G\boldsymbol{y} = D\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{y}) - D_1\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x}))\pi_1 HD\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{y})$$

As just explained, $L_{\boldsymbol{x}} \circ H$ is the identity on $P_{\boldsymbol{x}}$, the image of $D\boldsymbol{f}(\boldsymbol{x})$. Then

$$\begin{aligned} D_2\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x}))\pi_2 G\boldsymbol{y} &= L_{\boldsymbol{x}} \circ HD\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{y}) - D_1\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x}))\pi_1 HD\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{y}) \\ &= \left(L_{\boldsymbol{x}} \circ \underline{HD\boldsymbol{f}(\boldsymbol{x})} - D_1\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x}))\pi_1 \underline{HD\boldsymbol{f}(\boldsymbol{x})}\right)(\boldsymbol{y}) \end{aligned}$$

Factoring out that underlined term,

$$D_2\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x}))\pi_2 G\boldsymbol{y} = [L_{\boldsymbol{x}} - D_1\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x}))\pi_1]HD\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{y})$$

Now $D\boldsymbol{f}(\boldsymbol{x}) : M \to P_{\boldsymbol{x}} = D\boldsymbol{f}(\boldsymbol{x})(\mathbb{R}^n)$ is onto. (This is based on the assumption that $D\boldsymbol{f}(\boldsymbol{x})$ has rank m.) Thus it suffices to consider only $\boldsymbol{y} \in M$ in the right side of the above. However, for such $\boldsymbol{y}$,$\pi_2 G\boldsymbol{y} = \boldsymbol{0}$ because to be in $M, \psi_k(\boldsymbol{y}) = 0$ if $k \geq m+1$, and so the left side of the above equals **0.** Thus it appears this term on the left is **0 for any $\boldsymbol{y}$ chosen.** How can this be so? It can only take place if $D_2\boldsymbol{f}_1(\boldsymbol{g}(\boldsymbol{x})) = \boldsymbol{0}$ for every $\boldsymbol{x} \in V$. Thus, since $\boldsymbol{g}$ is onto, it can only take place if $D_2\boldsymbol{f}_1(\boldsymbol{x}) = \boldsymbol{0}$ for all $\boldsymbol{x} \in U$. Therefore on U it must be the case that $\boldsymbol{f}_1$ depends only on $x_1, \cdots, x_m$ as desired. ■

8.9 The Local Structure of C^1 Mappings

In linear algebra it is shown that every invertible matrix can be written as a product of elementary matrices, those matrices which are obtained from doing a row operation to the identity matrix. Two of the row operations produce a matrix which will change exactly one entry of a vector when it is multiplied by the elementary matrix. The other row operation involves switching two rows and this has the effect of switching two entries in a vector when multiplied on the left by the elementary matrix. Thus, in terms of the effect on a vector, the mapping determined by the given matrix can be considered as a composition of mappings which either flip two entries of the vector or change exactly one. A similar local result is available for nonlinear mappings. I found this interesting result in the advanced calculus book by Rudin.

Definition 8.9.1 *Let U be an open set in $\mathbb{R}^n$ and let $\boldsymbol{G} : U \to \mathbb{R}^n$. Then $\boldsymbol{G}$ is called primitive if it is of the form*

$$\boldsymbol{G}(\boldsymbol{x}) = \begin{pmatrix} x_1 & \cdots & \alpha(\boldsymbol{x}) & \cdots & x_n \end{pmatrix}^T.$$

Thus, $\boldsymbol{G}$ is primitive if it only changes one of the variables. A function $\boldsymbol{F} : \mathbb{R}^n \to \mathbb{R}^n$ is called a flip if

$$\boldsymbol{F}(x_1, \cdots, x_k, \cdots, x_l, \cdots, x_n) = (x_1, \cdots, x_l, \cdots, x_k, \cdots, x_n)^T.$$

Thus a function is a flip if it interchanges two coordinates. Also, for $m = 1, 2, \cdots, n$, define

$$P_m(\boldsymbol{x}) \equiv \begin{pmatrix} x_1 & x_2 & \cdots & x_m & 0 & \cdots & 0 \end{pmatrix}^T$$

It turns out that if $\boldsymbol{h}(\mathbf{0}) = \mathbf{0}, D\boldsymbol{h}(0)^{-1}$ exists, and $\boldsymbol{h}$ is C^1 on U, then $\boldsymbol{h}$ can be written as a composition of primitive functions and flips. This is a very interesting application of the inverse function theorem.

Theorem 8.9.2 *Let $\boldsymbol{h} : U \to \mathbb{R}^n$ be a C^1 function with $\boldsymbol{h}(\mathbf{0}) = \mathbf{0}, D\boldsymbol{h}(\mathbf{0})^{-1}$ exists. Then there is an open set $V \subseteq U$ containing $\mathbf{0}$, flips $\boldsymbol{F}_1, \cdots, \boldsymbol{F}_{n-1}$, and primitive functions $\boldsymbol{G}_n, \boldsymbol{G}_{n-1}, \cdots, \boldsymbol{G}_1$ such that for $\boldsymbol{x} \in V$,*

$$\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{F}_1 \circ \cdots \circ \boldsymbol{F}_{n-1} \circ \boldsymbol{G}_n \circ \boldsymbol{G}_{n-1} \circ \cdots \circ \boldsymbol{G}_1(\boldsymbol{x}).$$

The primitive function $\boldsymbol{G}_j$ leaves x_i unchanged for $i \neq j$.

Proof: Let

$$\boldsymbol{h}_1(\boldsymbol{x}) \equiv \boldsymbol{h}(\boldsymbol{x}) = \begin{pmatrix} \alpha_1(\boldsymbol{x}) & \cdots & \alpha_n(\boldsymbol{x}) \end{pmatrix}^T$$

$$D\boldsymbol{h}(\mathbf{0})\boldsymbol{e}_1 = \begin{pmatrix} \alpha_{1,1}(\mathbf{0}) & \cdots & \alpha_{n,1}(\mathbf{0}) \end{pmatrix}^T$$

where $\alpha_{k,1}$ denotes $\frac{\partial \alpha_k}{\partial x_1}$. Since $D\boldsymbol{h}(\mathbf{0})$ is one to one, the right side of this expression cannot be zero. Hence there exists some k such that $\alpha_{k,1}(\mathbf{0}) \neq 0$. Now define

$$\boldsymbol{G}_1(\boldsymbol{x}) \equiv \begin{pmatrix} \alpha_k(\boldsymbol{x}) & x_2 & \cdots & x_n \end{pmatrix}^T$$

Then the matrix of $D\boldsymbol{G}_1(\mathbf{0})$ is of the form

$$\begin{pmatrix} \alpha_{k,1}(\mathbf{0}) & \cdots & \cdots & \alpha_{k,n}(\mathbf{0}) \\ 0 & 1 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

and its determinant equals $\alpha_{k,1}(\mathbf{0}) \neq 0$. Therefore, by the inverse function theorem, there exists an open set U_1, containing $\mathbf{0}$ and an open set V_2 containing $\mathbf{0}$ such that $\boldsymbol{G}_1(U_1) = V_2$ and $\boldsymbol{G}_1$ is one to one and onto, such that it and its inverse are both C^1. Let $\boldsymbol{F}_1$ denote the flip which interchanges x_k with x_1. Now define

$$\boldsymbol{h}_2(\boldsymbol{y}) \equiv \boldsymbol{F}_1 \circ \boldsymbol{h}_1 \circ \boldsymbol{G}_1^{-1}(\boldsymbol{y})$$

Thus

$$\begin{aligned} \boldsymbol{h}_2(\boldsymbol{G}_1(\boldsymbol{x})) &\equiv \boldsymbol{F}_1 \circ \boldsymbol{h}_1(\boldsymbol{x}) \\ &= \begin{pmatrix} \alpha_k(\boldsymbol{x}) & \cdots & \alpha_1(\boldsymbol{x}) & \cdots & \alpha_n(\boldsymbol{x}) \end{pmatrix}^T \end{aligned} \tag{8.26}$$

Therefore,

$$P_1\boldsymbol{h}_2(\boldsymbol{G}_1(\boldsymbol{x})) = \begin{pmatrix} \alpha_k(\boldsymbol{x}) & 0 & \cdots & 0 \end{pmatrix}^T.$$

Also

$$P_1(\boldsymbol{G}_1(\boldsymbol{x})) = \begin{pmatrix} \alpha_k(\boldsymbol{x}) & 0 & \cdots & 0 \end{pmatrix}^T$$

so $P_1\boldsymbol{h}_2(\boldsymbol{y}) = P_1(\boldsymbol{y})$ for all $\boldsymbol{y} \in V_2$. Also, $\boldsymbol{h}_2(\mathbf{0}) = \mathbf{0}$ and $D\boldsymbol{h}_2(0)^{-1}$ exists because of the definition of $\boldsymbol{h}_2$ above and the chain rule. Since $\boldsymbol{F}_1^2 = I$, the identity map, it follows from (8.26) that

$$\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{h}_1(\boldsymbol{x}) = \boldsymbol{F}_1 \circ \boldsymbol{h}_2 \circ \boldsymbol{G}_1(\boldsymbol{x}). \tag{8.27}$$

Note that on an open set $V_2 \equiv G_1(U_1)$ containing the origin, $\boldsymbol{h}_2$ leaves the first entry unchanged. This is what $P_1\boldsymbol{h}_2(\boldsymbol{G}_1(\boldsymbol{x})) = P_1(\boldsymbol{G}_1(\boldsymbol{x}))$ says. In contrast, $\boldsymbol{h}_1 = \boldsymbol{h}$ left possibly no entries unchanged.

Suppose then, that for $m \geq 2$, $\boldsymbol{h}_m$ leaves the first $m-1$ entries unchanged,

$$P_{m-1}\boldsymbol{h}_m(\boldsymbol{x}) = P_{m-1}(\boldsymbol{x}) \tag{8.28}$$

for all $\boldsymbol{x} \in U_m$, an open subset of U containing $\mathbf{0}$, and $\boldsymbol{h}_m(\mathbf{0}) = \mathbf{0}$, $D\boldsymbol{h}_m(\mathbf{0})^{-1}$ exists. From (8.28), $\boldsymbol{h}_m(\boldsymbol{x})$ must be of the form

$$\boldsymbol{h}_m(\boldsymbol{x}) = \begin{pmatrix} x_1 & \cdots & x_{m-1} & \alpha_1(\boldsymbol{x}) & \cdots & \alpha_n(\boldsymbol{x}) \end{pmatrix}^T$$

where these α_k are different than the ones used earlier. Then

$$D\boldsymbol{h}_m(\mathbf{0})\boldsymbol{e}_m = \begin{pmatrix} 0 & \cdots & 0 & \alpha_{1,m}(\mathbf{0}) & \cdots & \alpha_{n,m}(\mathbf{0}) \end{pmatrix}^T \neq \mathbf{0}$$

because $D\boldsymbol{h}_m(\mathbf{0})^{-1}$ exists. Therefore, there exists a $k \geq m$ such that $\alpha_{k,m}(\mathbf{0}) \neq 0$, not the same k as before. Define

$$\boldsymbol{G}_m(\boldsymbol{x}) \equiv \begin{pmatrix} x_1 & \cdots & x_{m-1} & \alpha_k(\boldsymbol{x}) & \cdots & x_n \end{pmatrix}^T \tag{8.29}$$

so a change in $\boldsymbol{G}_m$ occurs only in the m^{th} slot. Then $G_m(\mathbf{0}) = \mathbf{0}$ and $D\boldsymbol{G}_m(\mathbf{0})^{-1}$ exists similar to the above. In fact

$$\det(D\boldsymbol{G}_m(\mathbf{0})) = \alpha_{k,m}(\mathbf{0}).$$

Therefore, by the inverse function theorem, there exists an open set V_{m+1} containing $\mathbf{0}$ such that $V_{m+1} = \boldsymbol{G}_m(U_m)$ with $\boldsymbol{G}_m$ and its inverse being one to one, continuous and onto. Let $\boldsymbol{F}_m$ be the flip which flips x_m and x_k. Then define $\boldsymbol{h}_{m+1}$ on V_{m+1} by

$$\boldsymbol{h}_{m+1}(\boldsymbol{y}) = \boldsymbol{F}_m \circ \boldsymbol{h}_m \circ \boldsymbol{G}_m^{-1}(\boldsymbol{y}).$$

Thus for $\boldsymbol{x} \in U_m$,

$$\boldsymbol{h}_{m+1}(\boldsymbol{G}_m(\boldsymbol{x})) = (\boldsymbol{F}_m \circ \boldsymbol{h}_m)(\boldsymbol{x}). \tag{8.30}$$

and consequently, since $\boldsymbol{F}_m^2 = I$,

$$\boldsymbol{F}_m \circ \boldsymbol{h}_{m+1} \circ \boldsymbol{G}_m(\boldsymbol{x}) = \boldsymbol{h}_m(\boldsymbol{x}) \tag{8.31}$$

It follows

$$\begin{aligned} P_m\boldsymbol{h}_{m+1}(\boldsymbol{G}_m(\boldsymbol{x})) &= P_m(\boldsymbol{F}_m \circ \boldsymbol{h}_m)(\boldsymbol{x}) \\ &= \begin{pmatrix} x_1 & \cdots & x_{m-1} & \alpha_k(\boldsymbol{x}) & 0 & \cdots & 0 \end{pmatrix}^T \end{aligned}$$

and

$$P_m(\boldsymbol{G}_m(\boldsymbol{x})) = \begin{pmatrix} x_1 & \cdots & x_{m-1} & \alpha_k(\boldsymbol{x}) & 0 & \cdots & 0 \end{pmatrix}^T.$$

Therefore, for $\boldsymbol{y} \in V_{m+1}$,

$$P_m\boldsymbol{h}_{m+1}(\boldsymbol{y}) = P_m(\boldsymbol{y}).$$

As before, $\boldsymbol{h}_{m+1}(\mathbf{0}) = \mathbf{0}$ and $D\boldsymbol{h}_{m+1}(\mathbf{0})^{-1}$ exists. Therefore, we can apply (8.31) repeatedly, obtaining the following:

$$\begin{aligned} \boldsymbol{h}(\boldsymbol{x}) &= \boldsymbol{F}_1 \circ \boldsymbol{h}_2 \circ \boldsymbol{G}_1(\boldsymbol{x}) \\ &= \boldsymbol{F}_1 \circ \boldsymbol{F}_2 \circ \boldsymbol{h}_3 \circ \boldsymbol{G}_2 \circ \boldsymbol{G}_1(\boldsymbol{x}) \\ &\vdots \\ &= \boldsymbol{F}_1 \circ \cdots \circ \boldsymbol{F}_{n-1} \circ \boldsymbol{h}_n \circ \boldsymbol{G}_{n-1} \circ \cdots \circ \boldsymbol{G}_1(\boldsymbol{x}) \end{aligned}$$

where $\boldsymbol{h}_n$ fixes the first $n-1$ entries,

$$P_{n-1}\boldsymbol{h}_n(\boldsymbol{x}) = P_{n-1}(\boldsymbol{x}) = \begin{pmatrix} x_1 & \cdots & x_{n-1} & 0 \end{pmatrix}^T,$$

and so $\boldsymbol{h}_n(\boldsymbol{x})$ is a primitive mapping of the form

$$\boldsymbol{h}_n(\boldsymbol{x}) = \begin{pmatrix} x_1 & \cdots & x_{n-1} & \alpha(\boldsymbol{x}) \end{pmatrix}^T.$$

Therefore, define the primitive function $\boldsymbol{G}_n(\boldsymbol{x})$ to equal $\boldsymbol{h}_n(\boldsymbol{x})$. ■

8.10 Invariance of Domain

As an application of the inverse function theorem is a simple proof of the important invariance of domain theorem which says that continuous and one to one functions defined on an open set in $\mathbb{R}^n$ with values in $\mathbb{R}^n$ take open sets to open sets. You know that this is true for functions of one variable because a one to one continuous function must be either strictly increasing or strictly decreasing. This will be used when considering orientations of curves later. However, the n dimensional version isn't at all obvious but is just as important if you want to consider manifolds with boundary for example. The need for this theorem occurs in many other places as well in addition to being extremely interesting for its own sake. The inverse function theorem gives conditions under which a differentiable function maps open sets to open sets. The following lemma, depending on the Brouwer fixed point theorem is the thing which will allow this to be extended to continuous one to one functions. It says roughly that if a continuous function does not move points near $\boldsymbol{p}$ very far, then the image of a ball centered at $\boldsymbol{p}$ contains an open set.

Lemma 8.10.1 *Let $\boldsymbol{f}$ be continuous and map $\overline{B(\boldsymbol{p},r)} \subseteq \mathbb{R}^n$ to $\mathbb{R}^n$. Suppose that for all $\boldsymbol{x} \in \overline{B(\boldsymbol{p},r)}$,*

$$|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{x}| < \varepsilon r$$

Then it follows that

$$\boldsymbol{f}\left(\overline{B(\boldsymbol{p},r)}\right) \supseteq B(\boldsymbol{p},(1-\varepsilon)r)$$

Proof: This is from the Brouwer fixed point theorem, Corollary 6.3.2. Consider for $\boldsymbol{y} \in B(\boldsymbol{p},(1-\varepsilon)r)$,

$$\boldsymbol{h}(\boldsymbol{x}) \equiv \boldsymbol{x} - \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{y}$$

Then $\boldsymbol{h}$ is continuous and for $\boldsymbol{x} \in \overline{B(\boldsymbol{p},r)}$,

$$|\boldsymbol{h}(\boldsymbol{x}) - \boldsymbol{p}| = |\boldsymbol{x} - \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{y} - \boldsymbol{p}| < \varepsilon r + |\boldsymbol{y} - \boldsymbol{p}| < \varepsilon r + (1-\varepsilon)r = r$$

Hence $\boldsymbol{h}:\overline{B(\boldsymbol{p},r)}\to\overline{B(\boldsymbol{p},r)}$ and so it has a fixed point $\boldsymbol{x}$ by Corollary 6.3.2 or Theorem 11.11.9. Thus

$$\boldsymbol{x}-\boldsymbol{f}(\boldsymbol{x})+\boldsymbol{y}=\boldsymbol{x}$$

so $\boldsymbol{f}(\boldsymbol{x})=\boldsymbol{y}$. ■

The notation $\|\boldsymbol{f}\|_K$ means $\sup_{\boldsymbol{x}\in K}|\boldsymbol{f}(\boldsymbol{x})|$. If you have a continuous function $\boldsymbol{h}$ defined on a compact set K, then the Stone Weierstrass theorem implies you can uniformly approximate it with a polynomial $\boldsymbol{g}$. That is $\|\boldsymbol{h}-\boldsymbol{g}\|_K$ is small. The following lemma says that you can also have $\boldsymbol{g}(\boldsymbol{z})=\boldsymbol{h}(\boldsymbol{z})$ and $D\boldsymbol{g}(\boldsymbol{z})^{-1}$ exists so that near $\boldsymbol{z}$, the function $\boldsymbol{g}$ will map open sets to open sets as claimed by the inverse function theorem. First is a little observation about approximating.

Lemma 8.10.2 *Suppose* $\det(A)=0$. *Then for all sufficiently small nonzero* ε,

$$\det(A+\varepsilon I)\neq 0$$

Proof: First suppose A is a $p\times p$ matrix. Suppose also that $\det(A)=0$. Thus, the constant term of $\det(\lambda I-A)$ is 0. Consider $\varepsilon I+A\equiv A_\varepsilon$ for small real ε. The characteristic polynomial of A_ε is

$$\det(\lambda I-A_\varepsilon)=\det((\lambda-\varepsilon)I-A)$$

This is of the form

$$(\lambda-\varepsilon)^p+a_{p-1}(\lambda-\varepsilon)^{p-1}+\cdots+(\lambda-\varepsilon)^m a_m$$

where the a_j are the coefficients in the characteristic equation for A and m is the largest such that $a_m\neq 0$. The constant term of this characteristic polynomial for A_ε must be nonzero for all ε small enough because it is of the form

$$(-1)^m\varepsilon^m a_m+(\text{higher order terms in }\varepsilon)$$

which shows that $\varepsilon I+A$ is invertible for all ε small enough but nonzero. ■

Lemma 8.10.3 *Let K be a compact set in $\mathbb{R}^n$ and let $\boldsymbol{h}:K\to\mathbb{R}^n$ be continuous, $\boldsymbol{z}\in K$ is fixed. Let $\delta>0$. Then there exists a polynomial $\boldsymbol{g}$ (each component a polynomial) such that*

$$\|\boldsymbol{g}-\boldsymbol{h}\|_K<\delta,\ \boldsymbol{g}(\boldsymbol{z})=\boldsymbol{h}(\boldsymbol{z}),\ D\boldsymbol{g}(\boldsymbol{z})^{-1}\ \text{exists}$$

Proof: By the Weierstrass approximation theorem, Corollary 5.7.8, or Theorem 5.9.5, there exists a polynomial $\hat{\boldsymbol{g}}$ such that

$$\|\hat{\boldsymbol{g}}-\boldsymbol{h}\|_K<\frac{\delta}{3}$$

Then define for $\boldsymbol{y}\in K$

$$\boldsymbol{g}(\boldsymbol{y})\equiv\hat{\boldsymbol{g}}(\boldsymbol{y})+\boldsymbol{h}(\boldsymbol{z})-\hat{\boldsymbol{g}}(\boldsymbol{z})$$

Then

$$\boldsymbol{g}(\boldsymbol{z})=\hat{\boldsymbol{g}}(\boldsymbol{z})+\boldsymbol{h}(\boldsymbol{z})-\hat{\boldsymbol{g}}(\boldsymbol{z})=\boldsymbol{h}(\boldsymbol{z})$$

Also

$$\begin{aligned}|\boldsymbol{g}(\boldsymbol{y})-\boldsymbol{h}(\boldsymbol{y})| &\leq |(\hat{\boldsymbol{g}}(\boldsymbol{y})+\boldsymbol{h}(\boldsymbol{z})-\hat{\boldsymbol{g}}(\boldsymbol{z}))-\boldsymbol{h}(\boldsymbol{y})|\\ &\leq |\hat{\boldsymbol{g}}(\boldsymbol{y})-\boldsymbol{h}(\boldsymbol{y})|+|\boldsymbol{h}(\boldsymbol{z})-\hat{\boldsymbol{g}}(\boldsymbol{z})|<\frac{2\delta}{3}\end{aligned}$$

and so since y was arbitrary,

$$\|g-h\|_K \leq \frac{2\delta}{3} < \delta$$

If $Dg(z)^{-1}$ exists, then this is what is wanted. If not, use Lemma 8.10.2 and note that for all η small enough, you could replace g with $y \to g(y)+\eta(y-z)$ and it will still be the case that $\|g-h\|_K < \delta$ along with $g(z)=h(z)$ but now $Dg(z)^{-1}$ exists. Simply use the modified g. ■

The main result is essentially the following lemma which combines the conclusions of the above.

Lemma 8.10.4 *Let $f : \overline{B(p,r)} \to \mathbb{R}^n$ where the ball is also in $\mathbb{R}^n$. Let f be one to one, f continuous. Then there exists $\delta > 0$ such that*

$$f\left(\overline{B(p,r)}\right) \supseteq B(f(p),\delta).$$

In other words, $f(p)$ is an interior point of $f\left(\overline{B(p,r)}\right)$.

Proof: Since $f\left(\overline{B(p,r)}\right)$ is compact, it follows that $f^{-1} : f\left(\overline{B(p,r)}\right) \to \overline{B(p,r)}$ is continuous. By Lemma 8.10.3, there exists a polynomial $g : f\left(\overline{B(p,r)}\right) \to \mathbb{R}^n$ such that

$$\begin{aligned} \left\|g-f^{-1}\right\|_{f\left(\overline{B(p,r)}\right)} &< \varepsilon r,\ \varepsilon<1,\ Dg(f(p))^{-1} \\ \text{exists, and } g(f(p)) &= f^{-1}(f(p)) = p \end{aligned}$$

From the first inequality in the above,

$$|g(f(x))-x| = \left|g(f(x))-f^{-1}(f(x))\right| \leq \left\|g-f^{-1}\right\|_{f\left(\overline{B(p,r)}\right)} < \varepsilon r$$

By Lemma 8.10.1,

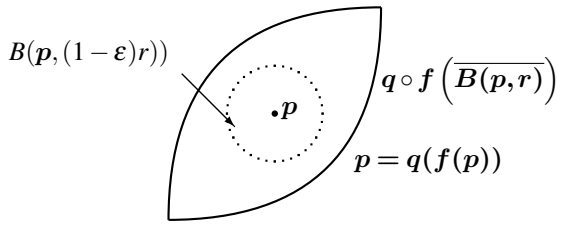
$$g\circ f\left(\overline{B(p,r)}\right) \supseteq B(p,(1-\varepsilon)r) = B(g(f(p)),(1-\varepsilon)r)$$

Since $Dg(f(p))^{-1}$ exists, it follows from the inverse function theorem that g^{-1} also exists and that g, g^{-1} are open maps on small open sets containing $f(p)$ and p respectively. Thus there exists $\eta < (1-\varepsilon)r$ such that g^{-1} is an open map on $B(p,\eta) \subseteq B(p,(1-\varepsilon)r)$. Thus

$$g\circ f\left(\overline{B(p,r)}\right) \supseteq B(p,(1-\varepsilon)r) \supseteq B(p,\eta)$$

So do g^{-1}‘ to both ends. Then you have $g^{-1}(p) = f(p)$ is in the open set $g^{-1}(B(p,\eta))$. Thus

$$f\left(\overline{B(p,r)}\right) \supseteq g^{-1}(B(p,\eta)) \supseteq B\left(g^{-1}(p),\delta\right) = B(f(p),\delta) \quad ■$$



With this lemma, the invariance of domain theorem comes right away. This remarkable theorem states that if $\boldsymbol{f} : U \to \mathbb{R}^n$ for U an open set in $\mathbb{R}^n$ and if $\boldsymbol{f}$ is one to one and continuous, then $\boldsymbol{f}(U)$ is also an open set in $\mathbb{R}^n$.

Theorem 8.10.5 *Let U be an open set in $\mathbb{R}^n$ and let $\boldsymbol{f} : U \to \mathbb{R}^n$ be one to one and continuous. Then $\boldsymbol{f}(U)$ is also an open subset in $\mathbb{R}^n$.*

Proof: It suffices to show that if $\boldsymbol{p} \in U$ then $\boldsymbol{f}(\boldsymbol{p})$ is an interior point of $\boldsymbol{f}(U)$. Let $\overline{B(\boldsymbol{p},r)} \subseteq U$. By Lemma 8.10.4, $\boldsymbol{f}(U) \supseteq \boldsymbol{f}\left(\overline{B(\boldsymbol{p},r)}\right) \supseteq B(\boldsymbol{f}(\boldsymbol{p}),\delta)$ so $\boldsymbol{f}(\boldsymbol{p})$ is indeed an interior point of $\boldsymbol{f}(U)$. ■

The inverse mapping theorem assumed quite a bit about the mapping. In particular it assumed that the mapping had a continuous derivative. The following version of the inverse function theorem seems very interesting because it only needs an invertible derivative at a point.

Corollary 8.10.6 *Let U be an open set in $\mathbb{R}^p$ and let $\boldsymbol{f} : U \to \mathbb{R}^p$ be one to one and continuous. Then, $\boldsymbol{f}^{-1}$ is also continuous on the open set $\boldsymbol{f}(U)$. If $\boldsymbol{f}$ is differentiable at $\boldsymbol{x}_1 \in U$ and if $D\boldsymbol{f}(\boldsymbol{x}_1)^{-1}$ exists for $\boldsymbol{x}_1 \in U$, then it follows that $D\boldsymbol{f}^{-1}(\boldsymbol{f}(\boldsymbol{x}_1)) = D\boldsymbol{f}(\boldsymbol{x}_1)^{-1}$.*

Proof: $|\cdot|$ will be a norm on $\mathbb{R}^p$, whichever is desired. If you like, let it be the Euclidean norm. $\|\cdot\|$ will be the operator norm. The first part of the conclusion of this corollary is from invariance of domain.

From the assumption that $D\boldsymbol{f}(\boldsymbol{x}_1)$ and $D\boldsymbol{f}(\boldsymbol{x}_1)^{-1}$ exists,

$$\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}_1) = \boldsymbol{f}\left(\boldsymbol{f}^{-1}(\boldsymbol{y})\right) - \boldsymbol{f}(\boldsymbol{x}_1) = D\boldsymbol{f}(\boldsymbol{x}_1)\left(\boldsymbol{f}^{-1}(\boldsymbol{y}) - \boldsymbol{x}_1\right) + \boldsymbol{o}\left(\boldsymbol{f}^{-1}(\boldsymbol{y}) - \boldsymbol{x}_1\right)$$

Since $D\boldsymbol{f}(\boldsymbol{x}_1)^{-1}$ exists,

$$D\boldsymbol{f}(\boldsymbol{x}_1)^{-1}(\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}_1)) = \boldsymbol{f}^{-1}(\boldsymbol{y}) - \boldsymbol{x}_1 + \boldsymbol{o}\left(\boldsymbol{f}^{-1}(\boldsymbol{y}) - \boldsymbol{x}_1\right)$$

by continuity, if $|\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}_1)|$ is small enough, then $\left|\boldsymbol{f}^{-1}(\boldsymbol{y}) - \boldsymbol{x}_1\right|$ is small enough that in the above,

$$\left|\boldsymbol{o}\left(\boldsymbol{f}^{-1}(\boldsymbol{y}) - \boldsymbol{x}_1\right)\right| < \frac{1}{2}\left|\boldsymbol{f}^{-1}(\boldsymbol{y}) - \boldsymbol{x}_1\right|$$

Hence, if $|\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}_1)|$ is sufficiently small, then from the triangle inequality of the form $|p - q| \geq ||p| - |q||$,

$$\begin{aligned}
\left\|D\boldsymbol{f}(\boldsymbol{x}_1)^{-1}\right\| |(\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}_1))| &\geq \left|D\boldsymbol{f}(\boldsymbol{x}_1)^{-1}(\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}_1))\right| \\
&\geq \left|\boldsymbol{f}^{-1}(\boldsymbol{y}) - \boldsymbol{x}_1\right| - \frac{1}{2}\left|\boldsymbol{f}^{-1}(\boldsymbol{y}) - \boldsymbol{x}_1\right| \\
&= \frac{1}{2}\left|\boldsymbol{f}^{-1}(\boldsymbol{y}) - \boldsymbol{x}_1\right|
\end{aligned}$$

$$|\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}_1)| \geq \left\|D\boldsymbol{f}(\boldsymbol{x}_1)^{-1}\right\|^{-1} \frac{1}{2}\left|\boldsymbol{f}^{-1}(\boldsymbol{y}) - \boldsymbol{x}_1\right|$$

It follows that for $|\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}_1)|$ small enough,

$$\left|\frac{\boldsymbol{o}\left(\boldsymbol{f}^{-1}(\boldsymbol{y}) - \boldsymbol{x}_1\right)}{\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}_1)}\right| \leq \left|\frac{\boldsymbol{o}\left(\boldsymbol{f}^{-1}(\boldsymbol{y}) - \boldsymbol{x}_1\right)}{\boldsymbol{f}^{-1}(\boldsymbol{y}) - \boldsymbol{x}_1}\right| \frac{2}{\left\|D\boldsymbol{f}(\boldsymbol{x}_1)^{-1}\right\|^{-1}}$$

Then, using continuity of the inverse function again, it follows that if $|\boldsymbol{y}-\boldsymbol{f}(\boldsymbol{x}_1)|$ is possibly still smaller, then $\boldsymbol{f}^{-1}(\boldsymbol{y})-\boldsymbol{x}_1$ is sufficiently small that the right side of the above inequality is no larger than ε. Since ε is arbitrary, it follows

$$\boldsymbol{o}\left(\boldsymbol{f}^{-1}(\boldsymbol{y})-\boldsymbol{x}_1\right)=\boldsymbol{o}(\boldsymbol{y}-\boldsymbol{f}(\boldsymbol{x}_1))$$

Now from differentiability of $\boldsymbol{f}$ at $\boldsymbol{x}_1$,

$$\begin{aligned}\boldsymbol{y}-\boldsymbol{f}(\boldsymbol{x}_1) &= \boldsymbol{f}\left(\boldsymbol{f}^{-1}(\boldsymbol{y})\right)-\boldsymbol{f}(\boldsymbol{x}_1)=D\boldsymbol{f}(\boldsymbol{x}_1)\left(\boldsymbol{f}^{-1}(\boldsymbol{y})-\boldsymbol{x}_1\right)+\boldsymbol{o}\left(\boldsymbol{f}^{-1}(\boldsymbol{y})-\boldsymbol{x}_1\right)\\ &= D\boldsymbol{f}(\boldsymbol{x}_1)\left(\boldsymbol{f}^{-1}(\boldsymbol{y})-\boldsymbol{x}_1\right)+\boldsymbol{o}(\boldsymbol{y}-\boldsymbol{f}(\boldsymbol{x}_1))\\ &= D\boldsymbol{f}(\boldsymbol{x}_1)\left(\boldsymbol{f}^{-1}(\boldsymbol{y})-\boldsymbol{f}^{-1}(\boldsymbol{f}(\boldsymbol{x}_1))\right)+\boldsymbol{o}(\boldsymbol{y}-\boldsymbol{f}(\boldsymbol{x}_1))\end{aligned}$$

Therefore, solving for $\boldsymbol{f}^{-1}(\boldsymbol{y})-\boldsymbol{f}^{-1}(\boldsymbol{f}(\boldsymbol{x}_1))$,

$$\boldsymbol{f}^{-1}(\boldsymbol{y})-\boldsymbol{f}^{-1}(\boldsymbol{f}(\boldsymbol{x}_1))=D\boldsymbol{f}(\boldsymbol{x}_1)^{-1}(\boldsymbol{y}-\boldsymbol{f}(\boldsymbol{x}_1))+\boldsymbol{o}(\boldsymbol{y}-\boldsymbol{f}(\boldsymbol{x}_1))$$

From the definition of the derivative, this shows that $D\boldsymbol{f}^{-1}(\boldsymbol{f}(\boldsymbol{x}_1))=D\boldsymbol{f}(\boldsymbol{x}_1)^{-1}$. ■

8.11 Exercises

1. This problem was suggested to me by Matt Heiner. Earlier there was a problem in which two surfaces intersected at a point and this implied that in fact, they intersected in a smooth curve. Now suppose you have two spheres $x^2+y^2+z^2=1$ and $(x-2)^2+y^2+z^2=1$. These intersect at the single point $(1,0,0)$. Why does the implicit function theorem not imply that these surfaces intersect in a curve?

2. Maximize $2x+y$ subject to the condition that $\frac{x^2}{4}+\frac{y^2}{9}\leq 1$. **Hint:** You need to consider interior points and also the method of Lagrange multipliers for the points on the boundary of this ellipse.

3. Maximize $x+y$ subject to the condition that $x^2+\frac{y^2}{9}+z^2\leq 1$.

4. Find the points on $y^2x=16$ which are closest to $(0,0)$.

5. Use Lagrange multipliers to "solve" the following maximization problem. Maximize xy^2z^3 subject to the constraint $x+y+z=12$. Show that the Lagrange multiplier method works very well but gives an answer which is neither a maximum nor a minimum. **Hint:** Show there is no maximum by considering $y=12-5x, z=4x$ and then letting x be large.

6. Let $f(x,y,z)=x^2-2yx+2z^2-4z+2$. Identify all the points where $Df=0$. Then determine whether they are local minima local maxima or saddle points.

7. Let $f(x,y)=x^4-2x^2+2y^2+1$. Identify all the points where $Df=0$. Then determine whether they are local minima local maxima or saddle points.

8. Let $f(x,y,z)=-x^4+2x^2-y^2-2z^2-1$. Identify all the points where $Df=0$. Then determine whether they are local minima local maxima or saddle points.

9. Let $f : V \to \mathbb{R}$ where V is a finite dimensional normed vector space. Suppose f is convex which means $f(t\boldsymbol{x}+(1-t)\boldsymbol{y}) \leq tf(\boldsymbol{x})+(1-t)f(\boldsymbol{y})$ whenever $t \in [0,1]$. Suppose also that f is differentiable. Show then that for every $\boldsymbol{x},\boldsymbol{y} \in V$,

$$(Df(\boldsymbol{x})-Df(\boldsymbol{y}))(\boldsymbol{x}-\boldsymbol{y}) \geq 0.$$

Thus convex functions have monotone derivatives.

10. Suppose B is an open ball in X and $\boldsymbol{f} : B \to Y$ is differentiable. Suppose also there exists $L \in \mathscr{L}(X,Y)$ such that $\|D\boldsymbol{f}(\boldsymbol{x})-L\| < k$ for all $\boldsymbol{x} \in B$. Show that if $\boldsymbol{x}_1,\boldsymbol{x}_2 \in B$,

$$|\boldsymbol{f}(\boldsymbol{x}_1)-\boldsymbol{f}(\boldsymbol{x}_2)-L(\boldsymbol{x}_1-\boldsymbol{x}_2)| \leq k|\boldsymbol{x}_1-\boldsymbol{x}_2|.$$

Hint: Consider $T\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x})-L\boldsymbol{x}$ and argue $\|DT(\boldsymbol{x})\| < k$.

11. Let $\boldsymbol{f} : U \subseteq X \to Y$, $D\boldsymbol{f}(\boldsymbol{x})$ exists for all $\boldsymbol{x} \in U$, $B(\boldsymbol{x}_0,\delta) \subseteq U$, and there exists $L \in \mathscr{L}(X,Y)$, such that $L^{-1} \in \mathscr{L}(Y,X)$, and for all $\boldsymbol{x} \in B(\boldsymbol{x}_0,\delta)$

$$\|D\boldsymbol{f}(\boldsymbol{x})-L\| < \frac{r}{\|L^{-1}\|},\ r<1.$$

Show that there exists $\varepsilon > 0$ and an open subset of $B(\boldsymbol{x}_0,\delta)$ called V, such that $\boldsymbol{f} : V \to B(\boldsymbol{f}(\boldsymbol{x}_0),\varepsilon)$ is one to one and onto. Also $D\boldsymbol{f}^{-1}(\boldsymbol{y})$ exists for each $\boldsymbol{y} \in B(f(\boldsymbol{x}_0),\varepsilon)$ and is given by the formula $D\boldsymbol{f}^{-1}(\boldsymbol{y}) = \left[D\boldsymbol{f}\left(\boldsymbol{f}^{-1}(\boldsymbol{y})\right)\right]^{-1}$. **Hint:** Let

$$T_{\boldsymbol{y}}(\boldsymbol{x}) \equiv T(\boldsymbol{x},\boldsymbol{y}) \equiv \boldsymbol{x}-L^{-1}(\boldsymbol{f}(\boldsymbol{x})-\boldsymbol{y})$$

for $|\boldsymbol{y}-\boldsymbol{f}(\boldsymbol{x}_0)| < \frac{(1-r)\delta}{2\|L^{-1}\|}$, consider $\{T_{\boldsymbol{y}}^n(\boldsymbol{x}_0)\}$. This is a version of the inverse function theorem for $\boldsymbol{f}$ only differentiable, not C^1.

12. If $\boldsymbol{f}$ is one to one and C^1, and $D\boldsymbol{f}(\boldsymbol{x}_0)$ is invertible, then locally the function $\boldsymbol{f}$ is one to one. Explain why this is, maybe using the above problem. However, this is a strictly local result! Let $\boldsymbol{f} : \mathbb{R}^2 \to \mathbb{R}^2$ be given by

$$\boldsymbol{f}(x,y) = \begin{pmatrix} e^x\cos y \\ e^x \sin y \end{pmatrix}$$

This clearly is not one to one because if you replace y with $y+2\pi$, you get the same value. Now verify that $D\boldsymbol{f}(x,y)^{-1}$ exists for all (x,y).

13. Show every polynomial, $\sum_{|\boldsymbol{\alpha}|\leq k} d_{\boldsymbol{\alpha}}\boldsymbol{x}^{\boldsymbol{\alpha}}$ is C^k for every k. Show that if f is defined and continuous on a compact set K, then there is an infinitely differentiable function which is uniformly close to f on K.

14. Suppose $U \subseteq \mathbb{R}^2$ is an open set and $\boldsymbol{f} : U \to \mathbb{R}^3$ is C^1. Suppose $D\boldsymbol{f}(s_0,t_0)$ has rank two and

$$\boldsymbol{f}(s_0,t_0) = \begin{pmatrix} x_0 & y_0 & z_0 \end{pmatrix}^T.$$

Show that for (s,t) near (s_0,t_0), the points $\boldsymbol{f}(s,t)$ may be realized in one of the following forms.

$$\{(x,y,\phi(x,y)) : (x,y) \text{ near } (x_0,y_0)\},$$

$$\{(\phi(y,z),y,z) : (y,z) \text{ near } (y_0,z_0)\},$$

or

$$\{(x,\phi(x,z),z,) : (x,z) \text{ near } (x_0,z_0)\}.$$

This shows that parametrically defined surfaces can be obtained locally in a particularly simple form.

15. Minimize $\sum_{j=1}^n x_j$ subject to the constraint $\sum_{j=1}^n x_j^2 = a^2$. Your answer should be some function of a which you may assume is a positive number.

16. A curve is formed from the intersection of the plane, $2x+3y+z=3$ and the cylinder $x^2+y^2=4$. Find the point on this curve which is closest to $(0,0,0)$.

17. A curve is formed from the intersection of the plane, $2x+3y+z=3$ and the sphere $x^2+y^2+z^2=16$. Find the point on this curve which is closest to $(0,0,0)$.

18. Let $A=(A_{ij})$ be an $n\times n$ matrix which is symmetric. Thus $A_{ij}=A_{ji}$ and recall $(A\boldsymbol{x})_i = A_{ij}x_j$ where you sum over the repeated index. Show $\frac{\partial}{\partial x_i}(A_{ij}x_jx_i) = 2A_{ij}x_j$. Show that when you use the method of Lagrange multipliers to maximize the function, $A_{ij}x_jx_i$ subject to the constraint, $\sum_{j=1}^n x_j^2 = 1$, the value of λ which corresponds to the maximum value of this functions is such that $A_{ij}x_j = \lambda x_i$. Thus $A\boldsymbol{x} = \lambda\boldsymbol{x}$. Thus λ is an eigenvalue of the matrix A.

19. Let $x_1,\cdots,x_5$ be 5 positive numbers. Maximize their product subject to the constraint that

$$x_1+2x_2+3x_3+4x_4+5x_5 = 300.$$

20. Let $f(x_1,\cdots,x_n) = x_1^n x_2^{n-1}\cdots x_n^1$. Then f achieves a maximum on the set,

$$S \equiv \left\{\boldsymbol{x}\in\mathbb{R}^n : \sum_{i=1}^n ix_i = 1 \text{ and each } x_i \geq 0\right\}.$$

If $\boldsymbol{x}\in S$ is the point where this maximum is achieved, find x_1/x_n.

21. Maximize $\prod_{i=1}^n x_i^2$ subject to the constraint, $\sum_{i=1}^n x_i^2 = r^2$. Show the maximum is $\left(r^2/n\right)^n$. Now show from this that $\left(\prod_{i=1}^n x_i^2\right)^{1/n} \leq \frac{1}{n}\sum_{i=1}^n x_i^2$ and finally, conclude that if each number $x_i \geq 0$, then

$$\left(\prod_{i=1}^n x_i\right)^{1/n} \leq \frac{1}{n}\sum_{i=1}^n x_i$$

and there exist values of the x_i for which equality holds. This says the "geometric mean" is always smaller than the arithmetic mean.

22. Show that there exists a smooth solution $y=y(x)$ to the equation

$$xe^y + ye^x = 0$$

which is valid for x,y both near 0. Find $y'(x)$ at a point (x,y) near $(0,0)$. Then find $y''(x)$ for such (x,y). Can you find an explicit formula for $y(x)$?

23. The next few problems involve invariance of domain. Suppose U is a nonempty open set in $\mathbb{R}^n$, $\boldsymbol{f} : U \to \mathbb{R}^n$ is continuous, and suppose that for each $\boldsymbol{x} \in U$, there is a ball $B_{\boldsymbol{x}}$ containing $\boldsymbol{x}$ such that $\boldsymbol{f}$ is one to one on $B_{\boldsymbol{x}}$. That is, $\boldsymbol{f}$ is locally one to one. Show that $\boldsymbol{f}(U)$ is open.

24. $\uparrow$ In the situation of the above problem, suppose $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^n$ is locally one to one. Also suppose that $\lim_{|\boldsymbol{x}|\to\infty} |\boldsymbol{f}(\boldsymbol{x})| = \infty$. Show it follows that $\boldsymbol{f}(\mathbb{R}^n) = \mathbb{R}^n$. That is, $\boldsymbol{f}$ is onto. Show that this would not be true if $\boldsymbol{f}$ is only defined on a proper open set. Also show that this would not be true if the condition $\lim_{|\boldsymbol{x}|\to\infty} |\boldsymbol{f}(\boldsymbol{x})| = \infty$ does not hold. **Hint:** You might show that $\boldsymbol{f}(\mathbb{R}^n)$ is both open and closed and then use connectedness. To get an example in the second case, you might think of e^{x+iy}. It does not include $0 + i0$. Why not?

25. $\uparrow$ Show that if $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^n$ is C^1 and if $D\boldsymbol{f}(\boldsymbol{x})$ exists and is invertible for all $\boldsymbol{x} \in \mathbb{R}^n$, then $\boldsymbol{f}$ is locally one to one. Thus, from the above problem, if $\lim_{|\boldsymbol{x}|\to\infty} |\boldsymbol{f}(\boldsymbol{x})| = \infty$, then $\boldsymbol{f}$ is also onto. Now consider $\boldsymbol{f} : \mathbb{R}^2 \to \mathbb{R}^2$ given by

$$\boldsymbol{f}(x,y) = \begin{pmatrix} e^x \cos y \\ e^x \sin y \end{pmatrix}$$

Show that this does not map onto $\mathbb{R}^2$. In fact, it fails to hit $(0,0)$, but $D\boldsymbol{f}(x,y)$ is invertible for all (x,y). Show why it fails to satisfy the limit condition.

26. In Example 8.1.9 on Page 214, could you replace y with $\boldsymbol{y} \in \mathbb{R}^m$ and obtain a modified version of this example?

Part II

Integration

Chapter 9

Measures and Measurable Functions

The Lebesgue integral is much better than the Rieman integral. This has been known for over 100 years. It is **much easier** to generalize to many dimensions and it is much easier to use in applications. It is also this integral which is most important in probability. However, this integral is more abstract. This chapter will develop the abstract machinery for this integral.

The next definition describes what is meant by a σ algebra. This is the fundamental object which is studied in probability theory. The events come from a σ algebra of sets. Recall that $\mathscr{P}(\Omega)$ is the set of all subsets of the given set Ω. It may also be denoted by 2^{Ω} but I won't refer to it this way.

Definition 9.0.1 *$\mathscr{F} \subseteq \mathscr{P}(\Omega)$, the set of all subsets of Ω, is called a σ algebra if it contains $\emptyset, \Omega$, and is closed with respect to countable unions and complements. That is, if $\{A_n\}_{n=1}^{\infty}$ is countable and each $A_n \in \mathscr{F}$, then $\cup_{n=1}^{\infty} A_m \in \mathscr{F}$ also and if $A \in \mathscr{F}$, then $\Omega \setminus A \in \mathscr{F}$. It is clear that any intersection of σ algebras is a σ algebra. If $\mathscr{K} \subseteq \mathscr{P}(\Omega)$, $\sigma(\mathscr{K})$ is the smallest σ algebra which contains $\mathscr{K}$. In fact, the intersection of all σ algebras containing $\mathscr{K}$ is obviously a σ algebra so this intersection is $\sigma(\mathscr{K})$.*

If $\mathscr{F}$ is a σ algebra, then it is also closed with respect to countable intersections. Here is why. Let $\{F_k\}_{k=1}^{\infty} \subseteq \mathscr{F}$. Then $(\cap_k F_k)^C = \cup_k F_k^C \in \mathscr{F}$ and so $\cap_k F_k = \left((\cap_k F_k)^C\right)^C = \left(\cup_k F_k^C\right)^C \in \mathscr{F}$.

Example 9.0.2 *You could consider $\mathbb{N}$ and for your σ algebra, you could have $\mathscr{P}(\mathbb{N})$. This satisfies all the necessary requirements. Note that in fact, $\mathscr{P}(S)$ works for any S. However, useful examples are not typically the set of all subsets.*

A useful idea is the idea of distance from a point to a set.

Definition 9.0.3 *Let (X,d) be a metric space and let S be a nonempty set in X. Then*

$$\operatorname{dist}(x,S) \equiv \inf\{d(x,y) : y \in S\}.$$

Recall the following lemma. It is Lemma 3.12.1, listed here for convenience.

Lemma 9.0.4 *The function, $x \to \operatorname{dist}(x,S)$ is continuous and in fact satisfies*

$$|\operatorname{dist}(x,S) - \operatorname{dist}(y,S)| \leq d(x,y).$$

9.1 Simple Functions and Measurable Functions

Recall that a σ algebra is a collection of subsets of a set Ω which includes $\emptyset, \Omega$, and is closed with respect to countable unions and complements.

Definition 9.1.1 *Let $(\Omega, \mathscr{F})$ be a measurable space, one for which $\mathscr{F}$ is a σ algebra contained in $\mathscr{P}(\Omega)$. Let $f : \Omega \to X$ where X is a metric space. Then f is measurable means that $f^{-1}(U) \in \mathscr{F}$ whenever U is open.*

It is important to have a theorem about pointwise limits of measurable functions, those with the property that inverse images of open sets are measurable. The following is a fairly general such theorem which holds in the situations to be considered in this book.

Theorem 9.1.2 *Let $\{f_n\}$ be a sequence of measurable functions mapping Ω to the metric space (X,d) where$(\Omega,\mathscr{F})$ is a measureable space. Suppose the pointwise limit $f(\omega)=\lim_{n\to\infty} f_n(\omega)$ for all ω. Then f is also a measurable function.*

Proof: It is required to show $f^{-1}(U)$ is measurable for all U open. Let

$$V_m \equiv \left\{ x \in U : \operatorname{dist}\left(x, U^C\right) > \frac{1}{m} \right\}.$$

Thus, since dist is continuous, $V_m \subseteq \left\{ x \in U : \operatorname{dist}\left(x,U^C\right) \geq \frac{1}{m} \right\}$ and $V_m \subseteq \overline{V_m} \subseteq V_{m+1}$ and $\cup_m V_m = U$. Then since V_m is open, $f^{-1}(V_m) = \cup_{n=1}^{\infty} \cap_{k=n}^{\infty} f_k^{-1}(V_m)$ and so

$$\begin{aligned} f^{-1}(U) &= \cup_{m=1}^{\infty} f^{-1}(V_m) = \cup_{m=1}^{\infty} \cup_{n=1}^{\infty} \cap_{k=n}^{\infty} f_k^{-1}(V_m) \\ &\subseteq \cup_{m=1}^{\infty} f^{-1}\left(\overline{V_m}\right) = f^{-1}(U) \end{aligned}$$

which shows $f^{-1}(U)$ is measurable. ■

An important example of a metric space is of course $\mathbb{R}, \mathbb{C}, \mathbb{F}^n$, where $\mathbb{F}$ is either $\mathbb{R}$ or $\mathbb{C}$ and so forth. However, it is also very convenient to consider the metric space $(-\infty,\infty]$, the real line with ∞ tacked on at the end. This can be considered as a metric space in a very simple way.

$$\rho(x,y) = |\arctan(x) - \arctan(y)|$$

with the understanding that $\arctan(\infty) \equiv \pi/2$. It is easy to show that this metric restricted to $\mathbb{R}$ gives the same open sets on $\mathbb{R}$ as the usual metric given by $d(x,y)=|x-y|$ but in addition, allows the inclusion of that ideal point out at the end of the real line denoted as ∞. This is considered mainly because it makes the development of the theory easier. The open sets in $(-\infty,\infty]$ are described in the following lemma.

Lemma 9.1.3 *The open balls in $(-\infty,\infty]$ consist of sets of the form (a,b) for a,b real numbers and $(a,\infty]$. This is a separable metric space.*

Proof: If the center of the ball is a real number, then the ball will result in an interval (a,b) where a,b are real numbers. If the center of the ball is ∞, then the ball results in something of the form $(a,\infty]$. It is obvious that this is a separable metric space with the countable dense set being $\mathbb{Q}$ since every ball contains a rational number. ■

If you kept both $-\infty$ and ∞ with the obvious generalization that $\arctan(-\infty) \equiv -\frac{\pi}{2}$, then the resulting metric space would be a complete separable metric space. However, it is not convenient to include $-\infty$, so this won't be done. The reason is that it will be desired to make sense of things like $f+g$.

Then for functions which have values in $(-\infty,\infty]$ we have the following extremely useful description of what it means for a function to be measurable.

Lemma 9.1.4 *Let $f:\Omega\to(-\infty,\infty]$ where $\mathscr{F}$ is a σ algebra of subsets of Ω. Here $(-\infty,\infty]$ is the metric space just described with the metric given by*

$$\rho(x,y) = |\arctan(x) - \arctan(y)|.$$

Then the following are equivalent.

$$f^{-1}((d,\infty]) \in \mathscr{F}, \text{for all finite } d,$$

$$f^{-1}((-\infty,d)) \in \mathscr{F}, \text{for all finite } d,$$

$$f^{-1}([d,\infty]) \in \mathscr{F}, \text{for all finite } d,$$

$$f^{-1}((-\infty,d]) \in \mathscr{F}, \text{for all finite } d,$$

$$f^{-1}((a,b)) \in \mathscr{F} \text{ for all } a<b, -\infty<a<b<\infty.$$

Any of these equivalent conditions is equivalent to the function being measurable.

Proof: First note that the first and the third are equivalent. To see this, observe $f^{-1}([d,\infty]) = \cap_{n=1}^{\infty} f^{-1}((d-1/n,\infty])$, and so if the first condition holds, then so does the third. $f^{-1}((d,\infty]) = \cup_{n=1}^{\infty} f^{-1}([d+1/n,\infty])$, and so if the third condition holds, so does the first.

Similarly, the second and fourth conditions are equivalent. Now from the definition of inverse image, $f^{-1}((-\infty,d]) = (f^{-1}((d,\infty]))^C$ so the first and fourth conditions are equivalent. Thus the first four conditions are equivalent and if any of them hold, then for $-\infty<a<b<\infty$, $f^{-1}((a,b)) = f^{-1}((-\infty,b)) \cap f^{-1}((a,\infty]) \in \mathscr{F}$. Finally, if the last condition holds, $f^{-1}([d,\infty]) = \left(\cup_{k=1}^{\infty} f^{-1}((-k+d,d))\right)^C \in \mathscr{F}$ and so the third condition holds. Therefore, all five conditions are equivalent.

Since $(-\infty,\infty]$ is a separable metric space, it follows from Theorem 3.4.2 that every open set U is a countable union of open intervals $U = \cup_k I_k$ where I_k is of the form (a,b) or $(a,\infty]$ and, as just shown if any of the equivalent conditions holds, then $f^{-1}(U) = \cup_k f^{-1}(I_k) \in \mathscr{F}$. Conversely, if $f^{-1}(U) \in \mathscr{F}$ for any open set $U \in (-\infty,\infty]$, then in particular, $f^{-1}((a,b)) \in \mathscr{F}$ which is one of the equivalent conditions and so all the equivalent conditions hold. ■

Note that if f is continuous and g is measurable, then $f \circ g$ is always measurable. This is because, for U open, $(f\circ g)^{-1}(U) = g^{-1}\left(f^{-1}(U)\right) = g^{-1}(\text{open})$ which is measurable.

There is a fundamental theorem about the relationship of simple functions to measurable functions given in the next theorem.

Definition 9.1.5 *Let $E \in \mathscr{F}$ for $\mathscr{F}$ a σ algebra. Then*

$$\mathscr{X}_E(\omega) \equiv \begin{cases} 1 \text{ if } \omega \in E \\ 0 \text{ if } \omega \notin E \end{cases}$$

This is called the indicator function of the set E. Let $s : (\Omega,\mathscr{F}) \to \mathbb{R}$. Then s is a simple function if it is of the form

$$s(\omega) = \sum_{i=1}^{n} c_i \mathscr{X}_{E_i}(\omega)$$

where $E_i \in \mathscr{F}$ and $c_i \in \mathbb{R}$, the E_i being disjoint. Thus simple functions have finitely many values and are measurable. In the next theorem, it will also be assumed that each $c_i \geq 0$.

Each simple function is measurable. This is easily seen as follows. First of all, you can assume the c_i are distinct because if not, you could just replace those E_i which correspond to a single value with their union. Then if you have any open interval (a,b), $s^{-1}((a,b)) = \cup\{E_i : c_i \in (a,b)\}$ and this is measurable because it is the finite union of measurable sets.

Theorem 9.1.6 *Let $f \geq 0$ be measurable. Then there exists a sequence of nonnegative simple functions $\{s_n\}$ satisfying*

$$0 \leq s_n(\omega) \tag{9.1}$$

$$\cdots s_n(\omega) \leq s_{n+1}(\omega) \cdots$$

$$f(\omega) = \lim_{n\to\infty} s_n(\omega) \text{ for all } \omega \in \Omega. \tag{9.2}$$

If f is bounded, the convergence is actually uniform. Conversely, if f is nonnegative and is the pointwise limit of such simple functions, then f is measurable.

Proof: Letting $I \equiv \{\omega : f(\omega) = \infty\}$, define

$$t_n(\omega) = \sum_{k=0}^{2^n} \frac{k}{n} \mathcal{X}_{f^{-1}\left(\left[\frac{k}{n}, \frac{k+1}{n}\right)\right)}(\omega) + 2^n \mathcal{X}_I(\omega).$$

Then $t_n(\omega) \leq f(\omega)$ for all ω and $\lim_{n\to\infty} t_n(\omega) = f(\omega)$ for all ω. This is because $t_n(\omega) = 2^n$ for $\omega \in I$ and if $f(\omega) \in [0, \frac{2^n+1}{n})$, then

$$0 \leq f(\omega) - t_n(\omega) \leq \frac{1}{n}. \tag{9.3}$$

Thus whenever $\omega \notin I$, the above inequality will hold for all n large enough. Let

$$s_1 = t_1,\ s_2 = \max(t_1, t_2),\ s_3 = \max(t_1, t_2, t_3), \cdots.$$

Then the sequence $\{s_n\}$ satisfies 9.1-9.2. Also each s_n has finitely many values and is measurable. To see this, note that $s_n^{-1}((a,\infty]) = \cup_{k=1}^n t_k^{-1}((a,\infty]) \in \mathcal{F}$

To verify the last claim, note that in this case the term $2^n \mathcal{X}_I(\omega)$ is not present and for n large enough, $2^n/n$ is larger than all values of f. Therefore, for all n large enough, 9.3 holds for all ω. Thus the convergence is uniform.

The last claim follows right away from Theorem 9.1.2. ■

There is a more general theorem which applies to measurable functions which have values in a separable metric space. In this context, a simple function is one which is of the form $\sum_{k=1}^m x_k \mathcal{X}_{E_k}(\omega)$ where the E_k are disjoint measurable sets and the x_k are in X. I am abusing notation somewhat by using a sum. You can't add in a general metric space. The symbol means the function has value x_k on the set E_k. However, if X were a vector space, this notation would be a nice way to express what is meant.

Theorem 9.1.7 *Let $(\Omega, \mathcal{F})$ be a measurable space and let $f : \Omega \to X$ where (X,d) is a separable metric space. Then f is a measurable function if and only if there exists a sequence of simple functions, $\{f_n\}$ such that for each $\omega \in \Omega$ and $n \in \mathbb{N}$,*

$$d(f_n(\omega), f(\omega)) \geq d(f_{n+1}(\omega), f(\omega)) \tag{9.4}$$

and

$$\lim_{n\to\infty} d(f_n(\omega), f(\omega)) = 0. \tag{9.5}$$

Proof: Let $D = \{x_k\}_{k=1}^{\infty}$ be a countable dense subset of X. First suppose f is measurable. Then since in a metric space every closed set C is the countable intersection of open sets,

$$C = \cap_{k=1}^{\infty} \{x \in X : \text{dist}(x, C) < 1/k\}$$

it follows $f^{-1}(\text{closed set}) \in \mathscr{F}$. Now let $D_n = \{x_k\}_{k=1}^{n}$. Let

$$A_1 \equiv \left\{ \omega : d(x_1, f(\omega)) = \min_{k \le n} d(x_k, f(\omega)) \right\}$$

$$= \cap_{k=1}^{n} \{\omega : d(x_k, f(\omega)) - d(x_1, f(\omega)) \ge 0\}$$

That is, A_1 is those ω such that $f(\omega)$ is approximated best out of D_n by x_1. Why is this a measurable set? It is because $\omega \to d(x_k, f(\omega)) - d(x_1, f(\omega))$ is a real valued measurable function, being the composition of a continuous function, $y \to d(x_k, y) - d(x_1, y)$ and a measurable function, $\omega \to f(\omega)$. Next let

$$A_2 \equiv \left\{ \omega \notin A_1 : d(x_2, f(\omega)) = \min_{k \le n} d(x_k, f(\omega)) \right\}$$

and continue in this manner obtaining disjoint measurable sets, $\{A_k\}_{k=1}^{n}$ such that for $\omega \in A_k$ the best approximation to $f(\omega)$ from D_n is x_k. Then $f_n(\omega) \equiv \sum_{k=1}^{n} x_k \mathscr{X}_{A_k}(\omega)$. Note

$$d(f_{n+1}(\omega), f(\omega)) = \min_{k \le n+1} d(x_k, f(\omega)) \le \min_{k \le n} d(x_k, f(\omega)) = d(f_n(\omega), f(\omega))$$

and so this verifies 9.4. It remains to verify 9.5.

Let $\varepsilon > 0$ be given and pick $\omega \in \Omega$. Then there exists $x_n \in D$ such that $d(x_n, f(\omega)) < \varepsilon$. It follows from the construction that

$$d(f_n(\omega), f(\omega)) \le d(x_n, f(\omega)) < \varepsilon.$$

This proves the first half.

Conversely, suppose the existence of the sequence of simple functions as described above. Each f_n is a measurable function because $f_n^{-1}(U) = \cup \{A_k : x_k \in U\}$. Therefore, the conclusion that f is measurable follows from Theorem 9.1.2 on Page 242. ■

Another useful observation is that the set where a sequence of measurable functions converges is also a measurable set.

Proposition 9.1.8 *Let $\{f_n\}$ be measurable with values in a complete normed vector space. Let $A \equiv \{\omega : \{f_n(\omega)\}$ converges$\}$. Then A is measurable.*

Proof: The set A is the same as the set on which $\{f_n(\omega)\}$ is a Cauchy sequence. This set is

$$\cap_{n=1}^{\infty} \cup_{m=1}^{\infty} \cap_{p,q>m} \left[\|f_p(\omega) - f_q(\omega)\| < \frac{1}{n} \right]$$

which is a measurable set thanks to the measurability of each f_n. ■

9.2 Measures and their Properties

What is meant by a **measure.**

Definition 9.2.1 *Let $(\Omega, \mathscr{F})$ be a measurable space. Here $\mathscr{F}$ is a σ algebra of sets of Ω. Then $\mu : \mathscr{F} \to [0,\infty]$ is called a measure if whenever $\{F_i\}_{i=1}^{\infty}$ is a sequence of disjoint sets of $\mathscr{F}$, it follows that*

$$\mu\left(\cup_{i=1}^{\infty} F_i\right) = \sum_{i=1}^{\infty} \mu\left(E_i\right)$$

Note that the series could equal ∞. If $\mu(\Omega) < \infty$, then μ is called a finite measure. An important case is when $\mu(\Omega) = 1$ when it is called a probability measure.

Note that $\mu(\emptyset) = \mu(\emptyset \cup \emptyset) = \mu(\emptyset) + \mu(\emptyset)$ and so $\mu(\emptyset) = 0$.

Example 9.2.2 *You could have $\mathscr{P}(\mathbb{N}) = \mathscr{F}$ and you could define $\mu(S)$ to be the number of elements of S. This is called counting measure. It is left as an exercise to show that this is a measure.*

Example 9.2.3 *Here is a pathological example. Let Ω be uncountable and $\mathscr{F}$ will be those sets which have the property that either the set is countable or its complement is countable. Let $\mu(E) = 0$ if E is countable and $\mu(E) = 1$ if E is uncountable. It is left as an exercise to show that this is a measure.*

Of course the most important measure in this book will be Lebesgue measure which gives the "volume" of a subset of $\mathbb{R}^n$. However, this requires a lot more work.

Lemma 9.2.4 *If μ is a measure and $F_i \in \mathscr{F}$, then $\mu\left(\cup_{i=1}^{\infty} F_i\right) \leq \sum_{i=1}^{\infty} \mu\left(F_i\right)$. Also if $F_n \in \mathscr{F}$ and $F_n \subseteq F_{n+1}$ for all n, then if $F = \cup_n F_n$,*

$$\mu(F) = \lim_{n\to\infty} \mu\left(F_n\right)$$

If $F_n \supseteq F_{n+1}$ for all n, then if $\mu(F_1) < \infty$ and $F = \cap_n F_n$, then

$$\mu(F) = \lim_{n\to\infty} \mu\left(F_n\right)$$

Proof: Let $G_1 = F_1$ and if $G_1, \cdots, G_n$ have been chosen disjoint, let

$$G_{n+1} \equiv F_{n+1} \setminus \cup_{i=1}^{n} G_i$$

Thus the G_i are disjoint. In addition, these are all measurable sets. Now

$$\mu\left(G_{n+1}\right) + \mu\left(F_{n+1} \cap \left(\cup_{i=1}^{n} G_i\right)\right) = \mu\left(F_{n+1}\right)$$

and so $\mu\left(G_n\right) \leq \mu\left(F_n\right)$. Therefore,

$$\mu\left(\cup_{i=1}^{\infty} G_i\right) = \sum_i \mu\left(G_i\right) \leq \sum_i \mu\left(F_i\right).$$

Now consider the increasing sequence of $F_n \in \mathscr{F}$. If $F \subseteq G$ and these are sets of $\mathscr{F}$

$$\mu(G) = \mu(F) + \mu(G \setminus F)$$

so $\mu(G) \geq \mu(F)$. Also

$$F = \cup_{i=1}^{\infty} (F_{i+1} \setminus F_i) + F_1$$

Then $\mu(F) = \sum_{i=1}^{\infty} \mu(F_{i+1} \setminus F_i) + \mu(F_1)$. Now $\mu(F_{i+1} \setminus F_i) + \mu(F_i) = \mu(F_{i+1})$. If any $\mu(F_i) = \infty$, there is nothing to prove. Assume then that these are all finite. Then

$$\mu(F_{i+1} \setminus F_i) = \mu(F_{i+1}) - \mu(F_i)$$

and so

$$\begin{aligned} \mu(F) &= \sum_{i=1}^{\infty} \mu(F_{i+1}) - \mu(F_i) + \mu(F_1) \\ &= \lim_{n\to\infty} \sum_{i=1}^{n} \mu(F_{i+1}) - \mu(F_i) + \mu(F_1) = \lim_{n\to\infty} \mu(F_{n+1}) \end{aligned}$$

Next suppose $\mu(F_1) < \infty$ and $\{F_n\}$ is a decreasing sequence. Then $F_1 \setminus F_n$ is increasing to $F_1 \setminus F$ and so by the first part,

$$\mu(F_1) - \mu(F) = \mu(F_1 \setminus F) = \lim_{n\to\infty} \mu(F_1 \setminus F_n) = \lim_{n\to\infty} (\mu(F_1) - \mu(F_n))$$

This is justified because $\mu(F_1 \setminus F_n) + \mu(F_n) = \mu(F_1)$ and all numbers are finite by assumption. Hence $\mu(F) = \lim_{n\to\infty} \mu(F_n)$. ■

I like to remember this as $E_n \uparrow E \Rightarrow \mu(E_n) \uparrow \mu(E)$ and $E_n \downarrow E \Rightarrow \mu(E_n) \downarrow \mu(E)$ if $\mu(E_1) < \infty$.

There is a monumentally important theorem called the Borel Cantelli lemma. This is next.

Lemma 9.2.5 *If $(\Omega, \mathscr{F}, \mu)$ is a measure space and if $\{E_i\} \subseteq \mathscr{F}$ and $\sum_{i=1}^{\infty} \mu(E_i) < \infty$, then there exists a set N of measure 0 ($\mu(N) = 0$) such that if $\omega \notin N$, then ω is in only finitely many of the E_i.*

Proof: The set of ω in infinitely many E_i is $N \equiv \cap_{n=1}^{\infty} \cup_{k\geq n} E_k$ because this consists of those ω which are in some E_k for $k \geq n$ for any choice of n. Now $\mu(N) \leq \sum_{k=n}^{\infty} \mu(E_k)$ which is just the tail of a convergent series. Thus, it converges to 0 as $n \to \infty$. Hence it is less than ε for n large enough. Thus $\mu(N)$ is no more than ε for any $\varepsilon > 0$. ■

9.3 Dynkin's Lemma

Dynkin's lemma is a very useful result. It is used quite a bit in books on probability. It will be used here to obtain n dimensional Lebesgue measure, also to establish an important result on regularity in the next section.

Definition 9.3.1 *Let Ω be a set and let $\mathscr{K}$ be a collection of subsets of Ω. Then $\mathscr{K}$ is called a π system if $\emptyset, \Omega \in \mathscr{K}$ and whenever $A, B \in \mathscr{K}$, it follows $A \cap B \in \mathscr{K}$.*

The following is the fundamental lemma which shows these π systems are useful. This is due to Dynkin.

Lemma 9.3.2 *Let $\mathscr{K}$ be a π system of subsets of Ω, a set. Also let $\mathscr{G}$ be a collection of subsets of Ω which satisfies the following three properties.*

1. $\mathscr{K} \subseteq \mathscr{G}$
2. *If* $A \in \mathscr{G}$, *then* $A^C \in \mathscr{G}$
3. *If* $\{A_i\}_{i=1}^{\infty}$ *is a sequence of disjoint sets from* $\mathscr{G}$ *then* $\cup_{i=1}^{\infty} A_i \in \mathscr{G}$.

Then $\mathscr{G} \supseteq \sigma(\mathscr{K})$, *where* $\sigma(\mathscr{K})$ *is the smallest* σ *algebra which contains* $\mathscr{K}$.

Proof: First note that if

$$\mathscr{H} \equiv \{\mathscr{G} : 1 \text{ - } 3 \text{ all hold}\}$$

then $\cap\mathscr{H}$ yields a collection of sets which also satisfies 1 - 3. Therefore, I will assume in the argument that $\mathscr{G}$ is the smallest collection satisfying 1 - 3. Let $A \in \mathscr{K}$ and define

$$\mathscr{G}_A \equiv \{B \in \mathscr{G} : A \cap B \in \mathscr{G}\}.$$

I want to show $\mathscr{G}_A$ satisfies 1 - 3 because then it must equal $\mathscr{G}$ since $\mathscr{G}$ is the smallest collection of subsets of Ω which satisfies 1 - 3. This will give the conclusion that for $A \in \mathscr{K}$ and $B \in \mathscr{G}$, $A \cap B \in \mathscr{G}$. This information will then be used to show that if $A, B \in \mathscr{G}$ then $A \cap B \in \mathscr{G}$. From this it will follow very easily that $\mathscr{G}$ is a σ algebra which will imply it contains $\sigma(\mathscr{K})$. Now here are the details of the argument.

Since $\mathscr{K}$ is given to be a π system contained in $\mathscr{G}$, $\mathscr{K} \subseteq \mathscr{G}_A$. Indeed, if $C \in \mathscr{K}$ then $A \cap C \in \mathscr{K} \subseteq \mathscr{G}$ so $C \in \mathscr{G}_A$. Property 3 is obvious because if $\{B_i\}$ is a sequence of disjoint sets in $\mathscr{G}_A$, then

$$A \cap \cup_{i=1}^{\infty} B_i = \cup_{i=1}^{\infty} A \cap B_i \in \mathscr{G}$$

because $A \cap B_i \in \mathscr{G}$ and the property 3 of $\mathscr{G}$.

It remains to verify Property 2 so let $B \in \mathscr{G}_A$. I need to verify that $B^C \in \mathscr{G}_A$. In other words, I need to show that $A \cap B^C \in \mathscr{G}$. However,

$$A \cap B^C = \left(A^C \cup (A \cap B)\right)^C \in \mathscr{G}$$

Here is why. Since $B \in \mathscr{G}_A$, $A \cap B \in \mathscr{G}$ and since $A \in \mathscr{K} \subseteq \mathscr{G}$ it follows $A^C \in \mathscr{G}$ by assumption 2. It follows from assumption 3 the union of the disjoint sets, A^C and $(A \cap B)$ is in $\mathscr{G}$ and then from 2 the complement of their union is in $\mathscr{G}$. Thus $\mathscr{G}_A$ satisfies 1 - 3 and this implies, since $\mathscr{G}$ is the smallest such, that $\mathscr{G}_A \supseteq \mathscr{G}$. However, $\mathscr{G}_A$ is constructed as a subset of $\mathscr{G}$. This proves that for every $B \in \mathscr{G}$ and $A \in \mathscr{K}$, $A \cap B \in \mathscr{G}$. Now pick $B \in \mathscr{G}$ and consider

$$\mathscr{G}_B \equiv \{A \in \mathscr{G} : A \cap B \in \mathscr{G}\}.$$

I just proved $\mathscr{K} \subseteq \mathscr{G}_B$. The other arguments are identical to show $\mathscr{G}_B$ satisfies 1 - 3 and is therefore equal to $\mathscr{G}$. This shows that whenever $A, B \in \mathscr{G}$ it follows $A \cap B \in \mathscr{G}$.

This implies $\mathscr{G}$ is a σ algebra. To show this, all that is left is to verify $\mathscr{G}$ is closed under countable unions because then it follows $\mathscr{G}$ is a σ algebra. Let $\{A_i\} \subseteq \mathscr{G}$. Then let $A_1' = A_1$ and

$$A_{n+1}' \equiv A_{n+1} \setminus (\cup_{i=1}^{n} A_i) = A_{n+1} \cap \left(\cap_{i=1}^{n} A_i^C\right) = \cap_{i=1}^{n} \left(A_{n+1} \cap A_i^C\right) \in \mathscr{G}$$

because finite intersections of sets of $\mathscr{G}$ are in $\mathscr{G}$. Since the A_i' are disjoint, it follows $\cup_{i=1}^{\infty} A_i = \cup_{i=1}^{\infty} A_i' \in \mathscr{G}$. Therefore, $\mathscr{G} \supseteq \sigma(\mathscr{K})$. ■

Corollary 9.3.3 *Given 2, closed with respect to complements, the condition that* $\mathscr{G}$ *is closed with respect to countable disjoint unions is equivalent to* $\mathscr{G}$ *the condition that* $\mathscr{G}$ *is closed with respect to countable intersections.*

Proof: $\Rightarrow$ Consider $\cap_k E_k$ where $E_k \in \mathscr{G}$. Then $\cap_k E_k = \left(\cup_k E_k^C\right)^C$. Now the E_k^C are not necessarily disjoint, but each is in $\mathscr{G}$ and so one can use the scheme of the last part of the proof of Lemma 9.3.2 to reduce to this case and conclude $\cup_k E_k^C \in \mathscr{G}$. Then the countable intersection is just the complement of this last set.

$\Leftarrow$ Suppose the countable intersection of sets of $\mathscr{G}$ is in $\mathscr{G}$ and consider a countable union $\cup_k E_k$ of sets of $\mathscr{G}$. Then $\cup_k E_k = \left(\cap_k E_k^C\right)^C \in \mathscr{G}$. ■

9.4 Measures and Regularity

It is often the case that Ω is not just a set. In particular, it is often the case that Ω is some sort of topological space, often a metric space. In this case, it is usually if not always the case that the open sets will be in the σ algebra of measurable sets. This leads to the following definition.

Definition 9.4.1 *A Polish space is a complete separable metric space. For a Polish space E or more generally a metric space or even a general topological space, $\mathscr{B}(E)$ denotes the **Borel sets** of E. This is defined to be the smallest σ algebra which contains the open sets. Thus it contains all open sets and closed sets and compact sets and many others.*

For example, $\mathbb{R}$ is a Polish space as is any separable Banach space. **Amazing things** can be said about finite measures on the Borel sets of a Polish space. First the case of a finite measure on a metric space will be considered.

It is best to not attempt to describe a generic Borel set. Always work with the definition that it is the smallest σ algebra containing the open sets. Attempts to give an explicit description of a "typical" Borel set tend to lead nowhere because there are so many things which can be done.You can take countable unions and complements and then countable intersections of what you get and then another countable union followed by complements and on and on. You just can't get a good useable description in this way. However, it is easy to see that something like $\left(\cap_{i=1}^{\infty}\cup_{j=i}^{\infty} E_j\right)^C$ is a Borel set if the E_j are. This is useful. This said, you can look at Hewitt and Stromberg in their discussion of why there are more Lebesgue measurable sets than Borel measurable sets to see the kind of technicalities which result by describing Borel sets. It is an extremely significant result based on describing Borel sets so it can be done.

Definition 9.4.2 *A measure μ defined on a σ algebra $\mathscr{F}$ which includes $\mathscr{B}(E)$ will be called inner regular on $\mathscr{F}$ if for all $F \in \mathscr{F}$,*

$$\mu(F) = \sup\{\mu(K) : K \subseteq F \text{ and } K \text{ is closed}\} \tag{9.6}$$

A measure, μ defined on $\mathscr{F}$ will be called outer regular on $\mathscr{F}$ if for all $F \in \mathscr{F}$,

$$\mu(F) = \inf\{\mu(V) : V \supseteq F \text{ and } V \text{ is open}\} \tag{9.7}$$

When a measure is both inner and outer regular, it is called regular. Actually, it is more useful and likely more standard to refer to μ being inner regular as

$$\mu(F) = \sup\{\mu(K) : K \subseteq F \text{ and } K \text{ is compact}\} \tag{9.8}$$

Thus the word "closed" is replaced with "compact". A complete measure defined on a σ algebra $\mathscr{F}$ which includes the Borel sets which is finite on compact sets and also satisfies

9.7 and 9.8 for each $F \in \mathscr{F}$ is called a Radon measure. A G_δ set, pronounced as G delta is the countable intersection of open sets. An F_σ set, pronounced F sigma is the countable union of closed sets.

In every case which has ever been of interest to me, the measure has been σ finite.

Definition 9.4.3 *If $(X, \mathscr{F}, \mu)$ is a measure space, it is called σ finite if there are $X_n \in \mathscr{F}$ with $\cup_n X_n = X$ and $\mu(X_n) < \infty$.*

For finite measures, defined on the Borel sets of a metric space X, $\mathscr{B}(X)$, the first definition of regularity is automatic. These are always outer and inner regular provided inner regularity refers to closed sets. Note that if $A \supseteq B$ then $A \setminus B = B^C \setminus A^C$.

Lemma 9.4.4 *Let μ be a finite measure defined on a σ algebra $\mathscr{F} \supseteq \mathscr{B}(X)$ where X is a metric space. Then the following hold.*

1. *μ is regular on $\mathscr{B}(X)$ meaning 9.6, 9.7 whenever $F \in \mathscr{B}(X)$.*
2. *μ is outer regular satisfying 9.7 on sets of $\mathscr{F}$ if and only if it is inner regular satisfying 9.6 on sets of $\mathscr{F}$.*
3. *If μ is either inner or outer regular on sets of $\mathscr{F}$ then if E is any set of $\mathscr{F}$, there exist F an F_σ set and G a G_δ set such that $F \subseteq E \subseteq G$ and $\mu(G \setminus F) = 0$.*

Proof: 1.) First note every open set is the countable union of closed sets and every closed set is the countable intersection of open sets. Here is why. Let V be an open set and let

$$K_k \equiv \left\{x \in V : \text{dist}\left(x, V^C\right) \geq 1/k\right\}.$$

Then clearly the union of the K_k equals V. Thus

$$\mu(V) = \sup\{\mu(K) : K \subseteq V \text{ and } K \text{ is closed}\}.$$

If U is open and contains V, then $\mu(U) \geq \mu(V)$ and so

$$\mu(V) \leq \inf\{\mu(U) : U \supseteq V,\ U \text{ open}\} \leq \mu(V) \text{ since } V \subseteq V.$$

Thus μ is inner and outer regular on open sets. In what follows, K will be closed and V will be open.

Let $\mathscr{K}$ be the open sets. This is a π system since it is closed with respect to finite intersections. Let

$$\mathscr{G} \equiv \{E \in \mathscr{B}(X) : \mu \text{ is inner and outer regular on } E\} \text{ so } \mathscr{G} \supseteq \mathscr{K}.$$

For $E \in \mathscr{G}$, let $V \supseteq E \supseteq K$ such that $\mu(V \setminus K) = \mu(V \setminus E) + \mu(E \setminus K) < \varepsilon$. Thus $K^C \supseteq E^C$ and so $\mu\left(K^C \setminus E^C\right) = \mu(E \setminus K) < \varepsilon$. Thus μ is outer regular on E^C because

$$\mu\left(K^C\right) = \mu\left(E^C\right) + \mu\left(K^C \setminus E^C\right) < \mu\left(E^C\right) + \varepsilon,\ K^C \supseteq E^C$$

Also, $E^C \supseteq V^C$ and $\mu\left(E^C \setminus V^C\right) = \mu(V \setminus E) < \varepsilon$ so μ is inner regular on E^C and so $\mathscr{G}$ is closed for complements. If the sets of $\mathscr{G}$ $\{E_i\}$ are disjoint, let $K_i \subseteq E_i \subseteq V_i$ with $\mu(V_i \setminus K_i) < \varepsilon 2^{-i}$. Then for $E \equiv \cup_i E_i$,and choosing m sufficiently large,

$$\mu(E) = \sum_i \mu(E_i) \leq \sum_{i=1}^m \mu(E_i) + \varepsilon \leq \sum_{i=1}^m \mu(K_i) + 2\varepsilon = \mu\left(\cup_{i=1}^m K_i\right) + 2\varepsilon$$

and so μ is inner regular on $E \equiv \cup_i E_i$. It remains to show that μ is outer regular on E. Letting $V \equiv \cup_i V_i$,

$$\mu(V \setminus E) \leq \mu(\cup_i (V_i \setminus E_i)) \leq \sum_i \varepsilon 2^{-i} = \varepsilon.$$

Hence μ is outer regular on E since $\mu(V) = \mu(E) + \mu(V \setminus E) \leq \mu(E) + \varepsilon$ and $V \supseteq E$.

By Dynkin's lemma, $\mathscr{G} = \sigma(\mathscr{K}) \equiv \mathscr{B}(X)$.

2.) Suppose that μ is outer regular on sets of $\mathscr{F} \supseteq \mathscr{B}(X)$. Letting $E \in \mathscr{F}$, by outer regularity, there exists an open set $V \supseteq E^C$ such that $\mu(V) - \mu(E^C) < \varepsilon$. Since μ is finite, $\varepsilon > \mu(V) - \mu(E^C) = \mu(V \setminus E^C) = \mu(E \setminus V^C) = \mu(E) - \mu(V^C)$ and V^C is a closed set contained in E. Therefore, if 9.7 holds, then so does 9.6. The converse is proved in the same way.

3.) The last claim is obtained by letting $G = \cap_n V_n$ where V_n is open, contains E, $V_n \supseteq V_{n+1}$, and $\mu(V_n) < \mu(E) + \frac{1}{n}$ and K_n, increasing closed sets contained in E such that $\mu(E) < \mu(K_n) + \frac{1}{n}$. Then let $F \equiv \cup F_n$ and $G \equiv \cap_n V_n$. Then $F \subseteq E \subseteq G$ and $\mu(G \setminus F) \leq \mu(V_n \setminus K_n) < 2/n$. ■

Next is a lemma which allows the replacement of closed with compact in the definition of inner regular.

Lemma 9.4.5 *Let μ be a finite measure on a σ algebra containing $\mathscr{B}(X)$, the Borel sets of X, a separable complete metric space, Polish space. Then if C is a closed set,*

$$\mu(C) = \sup\{\mu(K) : K \subseteq C \text{ and } K \text{ is compact.}\}$$

It follows that for a finite measure on $\mathscr{B}(X)$ where X is a Polish space, μ is inner regular in the sense that for all $F \in \mathscr{B}(X)$,

$$\mu(F) = \sup\{\mu(K) : K \subseteq F \text{ and } K \text{ is compact}\}$$

Proof: Let $\{a_k\}$ be a countable dense subset of C. Thus $\cup_{k=1}^{\infty} B\left(a_k, \frac{1}{n}\right) \supseteq C$. Therefore, there exists m_n such that

$$\mu\left(C \setminus \cup_{k=1}^{m_n} \overline{B\left(a_k, \frac{1}{n}\right)}\right) \equiv \mu(C \setminus C_n) < \frac{\varepsilon}{2^n}, \quad \cup_{k=1}^{m_n} \overline{B\left(a_k, \frac{1}{n}\right)} \equiv C_n.$$

Now let $K = C \cap (\cap_{n=1}^{\infty} C_n)$. Then K is a subset of C_n for each n and so for each $\varepsilon > 0$ there exists an ε net for K since C_n has a $1/n$ net, namely $a_1, \cdots, a_{m_n}$. Since K is closed, it is complete and so it is also compact since it is complete and totally bounded, Theorem 3.5.8. Now

$$\mu(C \setminus K) \leq \mu(\cup_{n=1}^{\infty} (C \setminus C_n)) < \sum_{n=1}^{\infty} \frac{\varepsilon}{2^n} = \varepsilon.$$

Thus $\mu(C)$ can be approximated by $\mu(K)$ for K a compact subset of C. The last claim follows from Lemma 9.4.4. ■

The next theorem is the main result. It says that if the measure is outer regular and μ is σ finite then there is an approximation for $E \in \mathscr{F}$ in terms of F_σ and G_δ sets in which the F_σ set is a countable union of compact sets. Also μ is inner and outer regular on $\mathscr{F}$.

Theorem 9.4.6 *Suppose $(X, \mathscr{F}, \mu)$, $\mathscr{F} \supseteq \mathscr{B}(X)$ is a measure space for X a metric space and μ is σ finite, $X = \cup_n X_n$ with $\mu(X_n) < \infty$ and the X_n disjoint. Suppose also that μ*

is outer regular. Then for each $E \in \mathscr{F}$, there exists F, G an F_σ and G_δ set respectively such that $F \subseteq E \subseteq G$ and $\mu(G \setminus F) = 0$. In particular, μ is inner and outer regular on $\mathscr{F}$. In case X is a complete separable metric space (Polish space), one can have F in the above be the countable union of compact sets and μ is inner regular in the sense of 9.8.

Proof: Since μ is outer regular and $\mu(X_n) < \infty$, there exists an open set $V_n \supseteq E \cap X_n$ such that

$$\mu(V_n \setminus (E \cap X_n)) = \mu(V_n) - \mu(E \cap X_n) < \frac{\varepsilon}{2^n}.$$

Then let $V \equiv \cup_n V_n$ so that $V \supseteq E$. Then $E = \cup_n E \cap X_n$ and so

$$\mu(V \setminus E) \leq \mu(\cup_n (V_n \setminus (E \cap X_n))) \leq \sum_n \mu(V_n \setminus (E \cap X_n)) < \sum_n \frac{\varepsilon}{2^n} = \varepsilon$$

Similarly, there exists U_n open such that $\mu\left(U_n \setminus \left(E^C \cap X_n\right)\right) < \frac{\varepsilon}{2^n}, U_n \supseteq E^C \cap X_n$ so if $U \equiv \cup_n U_n, \mu\left(U \setminus E^C\right) = \mu\left(E \setminus U^C\right) < \varepsilon$. Now U^C is closed and contained in E because $U \supseteq E^C$. Hence, letting $\varepsilon = \frac{1}{2^n}$, there exist closed sets C_n, and open sets V_n such that $C_n \subseteq E \subseteq V_n$ and $\mu(V_n \setminus C_n) < \frac{1}{2^{n-1}}$. Letting $G \equiv \cap_n V_n, F \equiv \cup_n C_n, F \subseteq E \subseteq G$ and $\mu(G \setminus F) \leq \mu(V_n \setminus C_n) < \frac{1}{2^{n-1}}$. Since n is arbitrary, $\mu(G \setminus F) = 0$.

To finish the proof, I will use Lemma 9.4.5 in the case where X is a Polish space.

By the first part, $\mu(G \setminus F) = 0$ where F is the countable union of closed sets $\{C_k\}_{k=1}^\infty$ and $F \subseteq E \subseteq G$. Letting $\mu_n(E) \equiv \mu(E \cap X_n)$, μ_n is a finite measure and so if C_k is one of those closed sets, Lemma 9.4.5 implies

$$\mu_n(C_k) \equiv \mu(C_k \cap X_n) = \sup\{\mu(K \cap X_n) : K \subseteq C_k,\ K \text{ compact}\}$$

Pick K_k compact such that $\mu_n(C_k \setminus K_k) < \frac{\varepsilon}{2^k}, K_k \subseteq C_k$. Then letting $\hat{F} \equiv \cup_k K_k$, it follows $\hat{F}$ is a countable union of compact sets contained in F and

$$\mu\left(F \setminus \hat{F}\right) = \mu(\cup_k C_k \setminus (\cup_k K_k)) \leq \mu(\cup_k (C_k \setminus K_k)) \leq \sum_k \mu(C_k \setminus K_k) < \varepsilon$$

Therefore, letting $\hat{F}_m$ be a countable union of compact sets contained in F for which $\mu\left(F \setminus \hat{F}_m\right) < \frac{1}{2^m}$, let $\tilde{F} \equiv \cup_m \hat{F}_m$. Then $\tilde{F}$ is a countable union of compact sets and

$$\mu\left(F \setminus \tilde{F}\right) \leq \mu\left(F \setminus \hat{F}_m\right) < \frac{1}{2^m}$$

and so $\mu\left(F \setminus \tilde{F}\right) = 0$. Then

$$\mu\left(G \setminus \tilde{F}\right) = \mu(G \setminus F) + \mu\left(F \setminus \tilde{F}\right) = \mu(G \setminus F) = 0$$

so as claimed, one can have F in the first part be the countable union of compact sets. Letting $E \in \mathscr{F}$, it was just shown that there exist G a G_δ set and F the countable union of compact sets such that $\mu(G \setminus F) = 0, F \subseteq E \subseteq G$. Therefore, $\mu(E) = \mu(E \setminus F) + \mu(F) = \mu(F)$ and so this shows inner regularity in the sense of 9.8 because if $l < \mu(E) = \mu(F)$, one could include enough of the compact sets whose union is F to obtain a compact set K for which $\mu(K) > l$. ■

An important example is the case of a random vector and its distribution measure.

Definition 9.4.7 *A measurable function $\boldsymbol{X}:(\Omega,\mathscr{F},\mu)\to Z$ a metric space is called a random variable when $\mu(\Omega)=1$. For such a random variable, one can define a distribution measure $\lambda_{\boldsymbol{X}}$ on the Borel sets of Z as follows. $\lambda_{\boldsymbol{X}}(G)\equiv\mu\left(\boldsymbol{X}^{-1}(G)\right)$. This is a well defined measure on the Borel sets of Z because it makes sense for every G open and $\mathscr{G}\equiv\left\{G\subseteq Z:\boldsymbol{X}^{-1}(G)\in\mathscr{F}\right\}$ is a σ algebra which contains the open sets, hence the Borel sets. Such a random variable is also called a random vector when Z is a vector space.*

Corollary 9.4.8 *Let $\boldsymbol{X}$ be a random variable with values in a separable complete metric space Z. Then $\lambda_{\boldsymbol{X}}$ is an inner and outer regular measure defined on $\mathscr{B}(Z)$.*

One such example of a complete metric space and a measure which is finite on compact sets is the following where the closures of balls are compact. Thus, this involves finite dimensional situations essentially. Note that if you have a metric space in which the closures of balls are compact sets, then the metric space must be separable. This is because you can pick a point ξ and consider the closures of balls $\overline{B(\xi,n)}$. Then $\overline{B(\xi,n)}$ is complete and totally bounded so it has a countable dense subset D_n. Let $D=\cup_n D_n$.

Corollary 9.4.9 *Let Ω be a complete metric space which is the countable union of compact sets K_n and suppose, for μ a Borel measure, $\mu(K_n)$ is finite. Then μ must be regular on $\mathscr{B}(\Omega)$. In particular, if Ω is a metric space and the closure of each ball is compact, and μ is finite on balls, then μ must be regular.*

Proof: Let the compact sets be increasing without loss of generality, and let $\mu_n(E)\equiv\mu(K_n\cap E)$. Thus μ_n is a finite measure defined on the Borel sets of a Polish space so it is regular. Letting $l<\mu(E)$, there exists n such that $l<\mu_n(E)\leq\mu(E)$. By what was shown above in Lemma 9.4.5, there exists H compact, $H\subseteq E$ such that also for a large n, $\mu_n(H)>l$. Hence $\mu(H\cap K_n)>l$ and so μ is inner regular. It remains to verify that μ is outer regular. If $\mu(E)=\infty$, there is nothing to show. Assume then that $\mu(E)<\infty$. Let $V_n\supseteq E$ with $\mu_n(V_n\setminus E)<\varepsilon 2^{-n}$ so also $\mu(V_n)<\infty$. We can assume also that $V_n\supseteq V_{n+1}$ for all n. Thus $\mu((V_n\setminus E)\cap K_n)<2^{-n}\varepsilon$. Let $G=\cap_k V_k$. Then $G\subseteq V_n$ so $\mu((G\setminus E)\cap K_n)<2^{-n}\varepsilon$. Letting $n\to\infty,\mu(G\setminus E)=0$ and $G\supseteq E$. Then, since V_1 has finite measure, $\mu(G\setminus E)=\lim_{n\to\infty}\mu(V_n\setminus E)$ and so for all n large enough, $\mu(V_n\setminus E)<\varepsilon$ so $\mu(E)+\varepsilon>\mu(V_n)$ and so μ is outer regular. In the last case, if the closure of each ball is compact, then Ω is automatically complete because every Cauchy sequence is contained in some ball and so has a convergent subsequence. Since the sequence is Cauchy, it also converges by Theorem 3.2.2 on Page 73. ■

9.5 Outer Measures

There is also something called an outer measure which is defined on the set of all subsets.

Definition 9.5.1 *Let Ω be a nonempty set and let $\lambda:\mathscr{P}(\Omega)\to[0,\infty)$ satisfy the following:*

1. $\lambda(\emptyset)=0$
2. *If $A\subseteq B$, then $\lambda(A)\leq\lambda(B)$*
3. $\lambda(\cup_{i=1}^{\infty}E_i)\leq\sum_{i=1}^{\infty}\lambda(E_i)$

Then λ is called an outer measure.

Every measure determines an outer measure. For example, suppose that μ is a measure on $\mathscr{F}$ a σ algebra of subsets of Ω. Then define

$$\bar{\mu}(S) \equiv \inf\{\mu(E) : E \supseteq S,\ E \in \mathscr{F}\}.$$

This is easily seen to be an outer measure. Also, we have the following Proposition.

Proposition 9.5.2 *Let μ be a measure defined on a σ algebra of subsets $\mathscr{F}$ of Ω as just described. Then $\bar{\mu}$ as defined above, is an outer measure and also, if $E \in \mathscr{F}$, then $\bar{\mu}(E) = \mu(E)$.*

Proof: The first two properties of an outer measure are obvious. What of the third? If any $\bar{\mu}(E_i) = \infty$, then there is nothing to show so suppose each of these is finite. Let $F_i \supseteq E_i$ such that $F_i \in \mathscr{F}$ and $\bar{\mu}(E_i) + \frac{\varepsilon}{2^i} > \mu(F_i)$. Then

$$\bar{\mu}(\cup_{i=1}^{\infty} E_i) \leq \mu(\cup_{i=1}^{\infty} F_i) \leq \sum_{i=1}^{\infty} \mu(F_i) < \sum_{i=1}^{\infty} \left(\bar{\mu}(E_i) + \frac{\varepsilon}{2^i}\right) = \sum_{i=1}^{\infty} \bar{\mu}(E_i) + \varepsilon$$

Since ε is arbitrary, this establishes the third condition. Finally, if $E \in \mathscr{F}$, then by definition, $\bar{\mu}(E) \leq \mu(E)$ because $E \supseteq E$. Also, $\mu(E) \leq \mu(F)$ for all $F \in \mathscr{F}$ such that $F \supseteq E$. It follows that $\mu(E)$ is a lower bound of all such $\mu(F)$ and so $\bar{\mu}(E) \geq \mu(E)$.■

9.6 Exercises

1. Show carefully that if $\mathfrak{S}$ is a set whose elements are σ algebras which are subsets of $\mathscr{P}(\Omega)$, then $\cap\mathfrak{S}$ is also a σ algebra. Now let $\mathscr{G} \subseteq \mathscr{P}(\Omega)$ satisfy property P if $\mathscr{G}$ is closed with respect to complements and countable disjoint unions as in Dynkin's lemma, and contains $\emptyset$ and Ω. If $\mathfrak{H} \subseteq \mathscr{G}$ is any set whose elements are subsets of $\mathscr{P}(\Omega)$ which satisfies property P, then $\cap\mathfrak{H}$ also satisfies property P. Thus there is a smallest subset of $\mathscr{G}$ satisfying P. In other words, verify the details of the proof of Dynkin's lemma.

2. The Borel sets of a metric space (X,d) are the sets in the smallest σ algebra which contains the open sets. These sets are denoted as $\mathscr{B}(X)$. Thus $\mathscr{B}(X) = \sigma(\text{open sets})$ where $\sigma(\mathscr{F})$ simply means the smallest σ algebra which contains $\mathscr{F}$. Show that in $\mathbb{R}^n$, $\mathscr{B}(\mathbb{R}^n) = \sigma(\mathscr{P})$ where $\mathscr{P}$ consists of the half open rectangles which are of the form $\prod_{i=1}^{n}[a_i, b_i)$.

3. Recall that $f : (\Omega, \mathscr{F}) \to X$ where X is a metric space is measurable means that inverse images of open sets are in $\mathscr{F}$. Show that if E is any set in $\mathscr{B}(X)$, then $f^{-1}(E) \in \mathscr{F}$. Thus, inverse images of Borel sets are measurable. Next consider $f : (\Omega, \mathscr{F}) \to X$ being measurable and $g : X \to Y$ is Borel measurable, meaning that $g^{-1}(open) \in \mathscr{B}(X)$. Explain why $g \circ f$ is measurable. **Hint:** You know that $(g \circ f)^{-1}(U) = f^{-1}\left(g^{-1}(U)\right)$. For your information, it does not work the other way around. That is, measurable composed with Borel measurable is not necessarily measurable. In fact examples exist which show that if g is measurable and f is continuous, then $g \circ f$ may fail to be measurable. An example is given later.

4. If you have X_i is a metric space, let $X = \prod_{i=1}^{n} X_i$ with the metric

$$d(\boldsymbol{x}, \boldsymbol{y}) \equiv \max\{d_i(x_i, y_i), i = 1, 2, \cdots, n\}$$

You considered this in an earlier problem. Show that any set of the form

$$\prod_{i=1}^{n} E_i,\ E_i \in \mathscr{B}(X_i)$$

is a Borel set. That is, the product of Borel sets is Borel. **Hint:** You might consider the continuous functions $\pi_i : \prod_{j=1}^{n} X_j \to X_i$ which are the projection maps. Thus $\pi_i(\boldsymbol{x}) \equiv x_i$. Then $\pi_i^{-1}(E_i)$ would have to be Borel measurable whenever $E_i \in \mathscr{B}(X_i)$. Explain why. You know π_i is continuous. Why would $\pi_i^{-1}(\textit{Borel})$ be a Borel set? Then you might argue that $\prod_{i=1}^{n} E_i = \cap_{i=1}^{n} \pi_i^{-1}(E_i)$.

5. You have two finite measures defined on $\mathscr{B}(X)$ μ, ν. Suppose these are equal on every open set. Show that these must be equal on every Borel set. **Hint:** You should use Dynkin's lemma to show this very easily.

6. Show that $(\mathbb{N}, \mathscr{P}(\mathbb{N}), \mu)$ is a measure space where $\mu(S)$ equals the number of elements of S. You need to verify that if the sets E_i are disjoint, then $\mu(\cup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} \mu(E_i)$.

7. Let Ω be an uncountable set and let $\mathscr{F}$ denote those subsets of Ω, F such that either F or F^C is countable. Show that this is a σ algebra. Next define the following measure. $\mu(A) = 1$ if A is uncountable and $\mu(A) = 0$ if A is countable. Show that μ is a measure. This is a perverted example.

8. Let $\mu(E) = 1$ if $0 \in E$ and $\mu(E) = 0$ if $0 \notin E$. Show this is a measure on $\mathscr{P}(\mathbb{R})$.

9. Give an example of a measure μ and a measure space and a decreasing sequence of measurable sets $\{E_i\}$ such that $\lim_{n\to\infty} \mu(E_n) \neq \mu(\cap_{i=1}^{\infty} E_i)$.

10. You have a measure space $(\Omega, \mathscr{F}, P)$ where P is a probability measure on $\mathscr{F}$. Then you also have a measurable function $X : \Omega \to Z$ where Z is some metric space. Thus $X^{-1}(U) \in \mathscr{F}$ whenever U is open. Now define a measure on $\mathscr{B}(Z)$ denoted by λ_X and defined by $\lambda_X(E) = P(\{\omega : X(\omega) \in E\})$. Explain why this yields a well defined probability measure on $\mathscr{B}(Z)$ which is regular. This is called the distribution measure.

11. Let $K \subseteq V$ where K is closed and V is open. Consider the following function.

$$f(x) = \frac{\operatorname{dist}(x, V^C)}{\operatorname{dist}(x, K) + \operatorname{dist}(x, V^C)}$$

Explain why this function is continuous, equals 0 off V and equals 1 on K.

12. Let $(\Omega, \mathscr{F})$ be a measurable space and let $f : \Omega \to X$ be a measurable function. Then $\sigma(f)$ denotes the smallest σ algebra such that f is measurable with respect to this σ algebra. Show that $\sigma(f) = \left\{f^{-1}(E) : E \in \mathscr{B}(X)\right\}$. More generally, you have a whole set of measurable functions $\mathscr{S}$ and $\sigma(\mathscr{S})$ denotes the smallest σ algebra such that each function in $\mathscr{S}$ is measurable. If you have an increasing list $\mathscr{S}_t$ for $t \in [0, \infty)$, then $\sigma(\mathscr{S}_t)$ will be what is called a filtration. You have a σ algebra for each $t \in [0, \infty)$ and as t increases, these σ algebras get larger. This is an essential part of the construction which is used to show that the Wiener process is a martingale. In fact the whole subject of martingales has to do with filtrations.

13. Let $(\Omega, \mathscr{F}, \mu)$ be a measure space. A sequence of functions $\{f_n\}$ is said to converge in measure to a measurable function f if and only if for each

$$\varepsilon > 0, \lim_{n\to\infty} \mu(\omega : |f_n(\omega) - f(\omega)| > \varepsilon) = 0.$$

Show that if this happens, then there exists a subsequence $\{f_{n_k}\}$ and a set of measure N such that if $\omega \notin N$, then $\lim_{k\to\infty} f_{n_k}(\omega) = f(\omega)$. Also show that if $\lim_{n\to\infty} f_n(\omega) = f(\omega)$, and $\mu(\Omega) < \infty$, then f_n converges in measure to f. **Hint:** For the subsequence, let $\mu\left(\omega : \left|f_{n_k}(\omega) - f(\omega)\right| > \varepsilon\right) < 2^{-k}$ and use Borel Cantelli lemma.

14. Let X, Y be separable metric spaces. Then $X \times Y$ can also be considered as a metric space with the metric $\rho((x,y),(\hat{x},\hat{y})) \equiv \max(d_X(x,\hat{x}), d_Y(y,\hat{y}))$. Verify this. Then show that if $\mathscr{K}$ consists of sets $A \times B$ where A, B are Borel sets in X and Y respectively, then it follows that $\sigma(\mathscr{K}) = \mathscr{B}(X \times Y)$, the Borel sets from $X \times Y$. Extend to the Cartesian product $\prod_i X_i$ of finitely many separable metric spaces.

9.7 An Outer Measure on $\mathscr{P}(\mathbb{R})$

A measure on $\mathbb{R}$ is like length. I will present something more general than length because it is no trouble to do so and the generalization is useful in many areas of mathematics such as probability.

Definition 9.7.1 *The following definition is important.*

$$F(x+) \equiv \lim_{y\to x+} F(y), \quad F(x-) = \lim_{y\to x-} F(y)$$

Thus one of these is the limit from the left and the other is the limit from the right.

In probability, one often has $F(x) \geq 0$, F is increasing, and $F(x+) = F(x)$. This is the case where F is a probability distribution function. In this case, $F(x) \equiv P(X \leq x)$ where X is a random variable. In this case, $\lim_{x\to\infty} F(x) = 1$ but we are considering more general functions than this including the simple example where $F(x) = x$. This last example will end up giving Lebesgue measure on $\mathbb{R}$. Recall the following definition.

Definition 9.7.2 *$\mathscr{P}(S)$ denotes the set of all subsets of S.*

Also recall

Definition 9.7.3 *For two sets, A, B in a metric space,*

$$\operatorname{dist}(A,B) \equiv \inf\{d(x,y) : x \in A, y \in B\}.$$

Theorem 9.7.4 *Let F be an increasing function defined on $\mathbb{R}$. This will be called an integrator function. There exists a function $\mu : \mathscr{P}(\mathbb{R}) \to [0,\infty]$ which satisfies the following properties.*

1. *If $A \subseteq B$, then $0 \leq \mu(A) \leq \mu(B), \mu(\emptyset) = 0$.*
2. $\mu\left(\cup_{k=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mu(A_i)$
3. $\mu([a,b]) = F(b+) - F(a-)$,

4. $\mu((a,b)) = F(b-) - F(a+)$

5. $\mu((a,b]) = F(b+) - F(a+)$

6. $\mu([a,b)) = F(b-) - F(a-)$.

7. *If* $\operatorname{dist}(A,B) = \delta > 0$, *then* $\mu(A \cup B) = \mu(A) + \mu(B)$.

Proof: First it is necessary to define the function μ. This is contained in the following definition.

Definition 9.7.5 *For* $A \subseteq \mathbb{R}$,

$$\mu(A) = \inf\left\{\sum_{i=1}^{\infty}(F(b_i-) - F(a_i+)) : A \subseteq \cup_{i=1}^{\infty}(a_i,b_i)\right\}$$

In words, you look at all coverings of A with open intervals. For each of these open coverings, you add the "lengths" of the individual open intervals and you take the infimum of all such numbers obtained.

Then 1.) is obvious because if a countable collection of open intervals covers B, then it also covers A. Thus the set of numbers obtained for B is smaller than the set of numbers for A. Why is $\mu(\emptyset) = 0$? Pick a point of continuity of F. Such points exist because F is increasing and so it has only countably many points of discontinuity. Let a be this point. Then $\emptyset \subseteq (a-\delta, a+\delta)$ and so $\mu(\emptyset) \leq F(a+\delta) - F(a-\delta)$ for every $\delta > 0$. Letting $\delta \to 0$, it follows that $\mu(\emptyset) = 0$.

Consider 2.). If any $\mu(A_i) = \infty$, there is nothing to prove. The assertion simply is $\infty \leq \infty$. Assume then that $\mu(A_i) < \infty$ for all i. Then for each $m \in \mathbb{N}$ there exists a countable set of open intervals, $\{(a_i^m, b_i^m)\}_{i=1}^{\infty}$ such that

$$\mu(A_m) + \frac{\varepsilon}{2^m} > \sum_{i=1}^{\infty}(F(b_i^m-) - F(a_i^m+)).$$

Then using Theorem 2.5.4 on Page 65,

$$\begin{aligned}\mu(\cup_{m=1}^{\infty}A_m) &\leq \sum_{i,m}(F(b_i^m-) - F(a_i^m+)) \\ &= \sum_{m=1}^{\infty}\sum_{i=1}^{\infty}(F(b_i^m-) - F(a_i^m+)) \leq \sum_{m=1}^{\infty}\mu(A_m) + \frac{\varepsilon}{2^m} = \sum_{m=1}^{\infty}\mu(A_m) + \varepsilon,\end{aligned}$$

and since ε is arbitrary, this establishes 2.).

Next consider 3.). By definition, there exists a sequence of open intervals, $\{(a_i,b_i)\}_{i=1}^{\infty}$ whose union contains $[a,b]$ such that

$$\mu([a,b]) + \varepsilon \geq \sum_{i=1}^{\infty}(F(b_i-) - F(a_i+)).$$

By Theorem 4.4.8, finitely many of these open intervals also cover $[a,b]$. It follows there exist finitely many of these intervals, denoted as $\{(a_i,b_i)\}_{i=1}^{n}$, which overlap, such that $a \in$

$(a_1,b_1), b_1 \in (a_2,b_2), \cdots, b \in (a_n,b_n)$. Therefore, $\mu([a,b]) \leq \sum_{i=1}^n (F(b_i-) - F(a_i+))$. It follows

$$\begin{aligned}\sum_{i=1}^n (F(b_i-) - F(a_i+)) &\geq \mu([a,b]) \geq \sum_{i=1}^n (F(b_i-) - F(a_i+)) - \varepsilon \\ &\geq F(b+) - F(a-) - \varepsilon\end{aligned}$$

Therefore, $F(b+\delta) - F(a-\delta) \geq \mu([a,b]) \geq F(b+) - F(a-) - \varepsilon$. Letting $\delta \to 0$,

$$F(b+) - F(a-) \geq \mu([a,b]) \geq F(b+) - F(a-) - \varepsilon$$

Since ε is arbitrary, this shows $\mu([a,b]) = F(b+) - F(a-)$. This establishes 3.).

Consider 4.). For small $\delta > 0, \mu([a+\delta, b-\delta]) \leq \mu((a,b))$ .Therefore, from 3.) and the definition of μ,

$$\begin{aligned}&F((b-\delta)) - F((a+\delta)) \leq F((b-\delta)+) - F((a+\delta)-) \\ &= \mu([a+\delta, b-\delta]) \leq \mu((a,b)) \leq F(b-) - F(a+)\end{aligned}$$

the last inequality from the definition. Now letting δ decrease to 0 it follows $F(b-) - F(a+) \leq \mu((a,b)) \leq F(b-) - F(a+)$. This shows 4.)

Consider 5.). From 3.) and 4.), for small $\delta > 0$,

$$\begin{aligned}&F(b+) - F((a+\delta)) \leq F(b+) - F((a+\delta)-) \\ = \;& \mu([a+\delta, b]) \leq \mu((a,b]) \leq \mu((a, b+\delta)) \\ = \;& F((b+\delta)-) - F(a+) \leq F(b+\delta) - F(a+).\end{aligned}$$

Now let δ converge to 0 from above to obtain $F(b+) - F(a+) = \mu((a,b])$. This establishes 5.) and 6.) is entirely similar to 5.).

Finally, consider 7.). Let

$$V \equiv \cup \left\{ B\left(x, \frac{\delta}{10}\right) : x \in A \cup B \right\}.$$

Let $A \cup B \subseteq \cup_{i=1}^\infty (a_i, b_i)$ where

$$\mu(A \cup B) + \varepsilon > \sum_i F(b_i-) - F(a_i+)$$

Then, taking the intersection of each of these intervals with V, it can be assumed that all of the intervals are contained in V since such an intersection will only strengthen the above inequality. Now refer to V as the union of these intervals, none of which can intersect both A and B. Thus V consists of disjoint open sets, one containing A consisting of the intervals which intersect A, U_A and the other consisting of those which intersect B, U_B. Let $\mathscr{I}_A$ denote the intervals which intersect A and let $\mathscr{I}_B$ denote the remaining intervals. Also let $\Delta((a_i,b_i)) \equiv F(b_i-) - F(a_i+)$. Then from the above,

$$\mu(A \cup B) + \varepsilon > \sum_{I \in \mathscr{I}_A} \Delta(I) + \sum_{I \in \mathscr{I}_B} \Delta(I) \geq \mu(A) + \mu(B) \geq \mu(A \cup B)$$

Since $\varepsilon > 0$ is arbitrary, this shows 7.). ■

We have just shown that μ is an outer measure on $\mathscr{P}(\mathbb{R})$. Unlike what was presented earlier, this outer measure did not begin with a measure.

9.8 Measures From Outer Measures

Theorem 9.7.4 describes an outer measure on $\mathscr{P}(\mathbb{R})$. There is a general procedure for constructing a σ algebra and a measure from an outer measure which is due to Caratheodory about 1918.

Thus, when you have a measure on $(\Omega, \mathscr{F})$, you can obtain an outer measure on $(\Omega, \mathscr{P}(\Omega))$ from this measure as in Proposition 9.5.2, and if you have an outer measure on $(\Omega, \mathscr{P}(\Omega))$, this will define a σ algebra $\mathscr{F}$ and a measure on $(\Omega, \mathscr{F})$. This last assertion is the topic of this section.

Definition 9.8.1 *Let Ω be a nonempty set and let $\mu : \mathscr{P}(\Omega) \to [0,\infty]$ be an outer measure. For $E \subseteq \Omega$, E is μ measurable if for all $S \subseteq \Omega$,*

$$\mu(S) = \mu(S \setminus E) + \mu(S \cap E). \tag{9.9}$$

To help in remembering 9.9, think of a measurable set E, as a process which divides a given set into two pieces, the part in E and the part not in E as in 9.9. In the Bible, there are several incidents recorded in which a process of division resulted in more stuff than was originally present.[1] Measurable sets are exactly those which are incapable of such a miracle. With an outer measure, it is always the case that $\mu(S) \leq \mu(S \setminus E) + \mu(S \cap E)$. The set is measurable, when equality is always obtained for any choice of $S \in \mathscr{P}(\Omega)$. You might think of the measurable sets as the non-miraculous sets. The idea is to show that these sets form a σ algebra on which the outer measure μ is a measure.

First here is a definition and a lemma.

Definition 9.8.2 *$(\mu \lfloor S)(A) \equiv \mu(S \cap A)$ for all $A \subseteq \Omega$. Thus $\mu \lfloor S$ is the name of a new outer measure, called μ restricted to S.*

The next lemma indicates that the property of measurability is not lost by considering this restricted measure.

Lemma 9.8.3 *If A is μ measurable, then for any S, A is $\mu \lfloor S$ measurable.*

Proof: Suppose A is μ measurable. It is desired to to show that for all $T \subseteq \Omega$,

$$(\mu \lfloor S)(T) = (\mu \lfloor S)(T \cap A) + (\mu \lfloor S)(T \setminus A).$$

Thus it is desired to show

$$\mu(S \cap T) = \mu(T \cap A \cap S) + \mu(T \cap S \cap A^C). \tag{9.10}$$

But 9.10 holds because A is μ measurable. Apply Definition 9.8.1 to $S \cap T$ instead of S. ■

If A is $\mu \lfloor S$ measurable, it does not follow that A is μ measurable. Indeed, if you believe in the existence of non measurable sets which is discussed later, you could let $A = S$ for such a μ non measurable set and verify that S is $\mu \lfloor S$ measurable.

The next theorem is the main result on outer measures which shows that starting with an outer measure you can obtain a measure.

[1] 1 Kings 17, 2 Kings 4, Mathew 14, and Mathew 15 all contain such descriptions. The stuff involved was either oil, bread, flour or fish. In mathematics such things have also been done with sets. In the book by Bruckner Bruckner and Thompson there is an interesting discussion of the Banach Tarski paradox which says it is possible to divide a ball in $\mathbb{R}^3$ into five disjoint pieces and assemble the pieces to form two disjoint balls of the same size as the first. The details can be found in: The Banach Tarski Paradox by Wagon, Cambridge University press. 1985. It is known that all such examples must involve the axiom of choice.

Theorem 9.8.4 *Let Ω be a set and let μ be an outer measure on $\mathscr{P}(\Omega)$. The collection of μ measurable sets $\mathscr{S}$, forms a σ algebra and*

$$\text{If } F_i \in \mathscr{S},\ F_i \cap F_j = \emptyset, \text{ then } \mu(\cup_{i=1}^{\infty} F_i) = \sum_{i=1}^{\infty} \mu(F_i). \tag{9.11}$$

If $\cdots F_n \subseteq F_{n+1} \subseteq \cdots$, then if $F = \cup_{n=1}^{\infty} F_n$ and $F_n \in \mathscr{S}$, it follows that

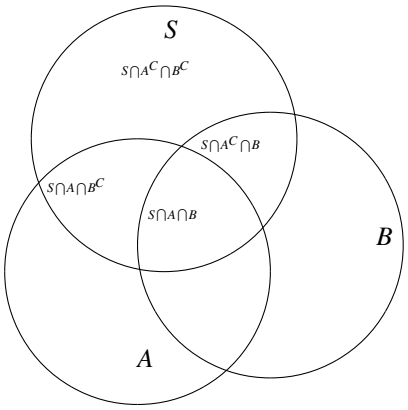
$$\mu(F) = \lim_{n\to\infty} \mu(F_n). \tag{9.12}$$

If $\cdots F_n \supseteq F_{n+1} \supseteq \cdots$, and if $F = \cap_{n=1}^{\infty} F_n$ for $F_n \in \mathscr{S}$ then if $\mu(F_1) < \infty$,

$$\mu(F) = \lim_{n\to\infty} \mu(F_n). \tag{9.13}$$

This measure space is also complete which means that if $\mu(F) = 0$ for some $F \in \mathscr{S}$ then if $G \subseteq F$, it follows $G \in \mathscr{S}$ also.

Proof: First note that $\emptyset$ and Ω are obviously in $\mathscr{S}$. Now suppose $A, B \in \mathscr{S}$. I will show $A \setminus B \equiv A \cap B^C$ is in $\mathscr{S}$. To do so, consider the following picture.



It is required to show that $\mu(S) = \mu(S \setminus (A \setminus B)) + \mu(S \cap (A \setminus B))$. First consider $S \setminus (A \setminus B)$. From the picture, it equals

$$\left(S \cap A^C \cap B^C\right) \cup (S \cap A \cap B) \cup \left(S \cap A^C \cap B\right)$$

Therefore, $\mu(S) \leq \mu(S \setminus (A \setminus B)) + \mu(S \cap (A \setminus B))$

$$\begin{aligned} &\leq \mu\left(S \cap A^C \cap B^C\right) + \mu(S \cap A \cap B) + \mu\left(S \cap A^C \cap B\right) + \mu(S \cap (A \setminus B)) \\ &= \mu\left(S \cap A^C \cap B^C\right) + \mu(S \cap A \cap B) + \mu\left(S \cap A^C \cap B\right) + \mu\left(S \cap A \cap B^C\right) \\ &= \mu\left(S \cap A^C \cap B^C\right) + \mu\left(S \cap A \cap B^C\right) + \mu(S \cap A \cap B) + \mu\left(S \cap A^C \cap B\right) \\ &= \mu\left(S \cap B^C\right) + \mu(S \cap B) = \mu(S) \end{aligned}$$

and so this shows that $A \setminus B \in \mathscr{S}$ whenever $A, B \in \mathscr{S}$.

Since $\Omega \in \mathcal{S}$, this shows that $A \in \mathcal{S}$ if and only if $A^C \in \mathcal{S}$. Now if $A, B \in \mathcal{S}$, $A \cup B = (A^C \cap B^C)^C = (A^C \setminus B)^C \in \mathcal{S}$. By induction, if $A_1, \cdots, A_n \in \mathcal{S}$, then so is $\cup_{i=1}^n A_i$. If $A, B \in \mathcal{S}$, with $A \cap B = \emptyset$,

$$\mu(A \cup B) = \mu((A \cup B) \cap A) + \mu((A \cup B) \setminus A) = \mu(A) + \mu(B).$$

By induction, if $A_i \cap A_j = \emptyset$ and $A_i \in \mathcal{S}$,

$$\mu(\cup_{i=1}^n A_i) = \sum_{i=1}^n \mu(A_i). \tag{9.14}$$

Now let $A = \cup_{i=1}^\infty A_i$ where $A_i \cap A_j = \emptyset$ for $i \neq j$. $\sum_{i=1}^\infty \mu(A_i) \geq \mu(A) \geq \mu(\cup_{i=1}^n A_i) = \sum_{i=1}^n \mu(A_i)$. Since this holds for all n, you can take the limit as $n \to \infty$ and conclude, $\sum_{i=1}^\infty \mu(A_i) = \mu(A)$ which establishes 9.11.

Consider part 9.12. Without loss of generality $\mu(F_k) < \infty$ for all k since otherwise there is nothing to show. Suppose $\{F_k\}$ is an increasing sequence of sets of $\mathcal{S}$. Then letting $F_0 \equiv \emptyset$, $\{F_{k+1} \setminus F_k\}_{k=0}^\infty$ is a sequence of disjoint sets of $\mathcal{S}$ since it was shown above that the difference of two sets of $\mathcal{S}$ is in $\mathcal{S}$. Also note that from 9.14

$$\mu(F_{k+1} \setminus F_k) + \mu(F_k) = \mu(F_{k+1})$$

and so if $\mu(F_k) < \infty$, then

$$\mu(F_{k+1} \setminus F_k) = \mu(F_{k+1}) - \mu(F_k).$$

Therefore, letting $F \equiv \cup_{k=1}^\infty F_k$ which also equals $\cup_{k=1}^\infty (F_{k+1} \setminus F_k)$, it follows from part 9.11 just shown that

$$\begin{aligned} \mu(F) &= \sum_{k=0}^\infty \mu(F_{k+1} \setminus F_k) = \lim_{n\to\infty} \sum_{k=0}^n \mu(F_{k+1} \setminus F_k) \\ &= \lim_{n\to\infty} \sum_{k=0}^n \mu(F_{k+1}) - \mu(F_k) = \lim_{n\to\infty} \mu(F_{n+1}). \end{aligned}$$

In order to establish 9.13, let the F_n be as given there. Then, since $(F_1 \setminus F_n)$ increases to $(F_1 \setminus F)$, 9.12 implies

$$\lim_{n\to\infty} (\mu(F_1) - \mu(F_n)) = \lim_{n\to\infty} \mu(F_1 \setminus F_n) = \mu(F_1 \setminus F).$$

The problem is, I don't know $F \in \mathcal{S}$ and so it is not clear that $\mu(F_1 \setminus F) = \mu(F_1) - \mu(F)$. However, $\mu(F_1 \setminus F) + \mu(F) \geq \mu(F_1)$ and so $\mu(F_1 \setminus F) \geq \mu(F_1) - \mu(F)$. Hence

$$\lim_{n\to\infty} (\mu(F_1) - \mu(F_n)) = \mu(F_1 \setminus F) \geq \mu(F_1) - \mu(F)$$

which implies $\lim_{n\to\infty} \mu(F_n) \leq \mu(F)$. But since $F \subseteq F_n$, $\mu(F) \leq \lim_{n\to\infty} \mu(F_n)$ and this establishes 9.13. Note that it was assumed $\mu(F_1) < \infty$ because $\mu(F_1)$ was subtracted from both sides.

It remains to show $\mathcal{S}$ is closed under countable unions. Recall that if $A \in \mathcal{S}$, then $A^C \in \mathcal{S}$ and $\mathcal{S}$ is closed under finite unions. Let $A_i \in \mathcal{S}$, $A = \cup_{i=1}^\infty A_i$, $B_n = \cup_{i=1}^n A_i$. Then

$$\begin{aligned} \mu(S) &= \mu(S \cap B_n) + \mu(S \setminus B_n) \\ &= (\mu \lfloor S)(B_n) + (\mu \lfloor S)(B_n^C). \end{aligned} \tag{9.15}$$

By Lemma 9.8.3 B_n is $(\mu\lfloor S)$ measurable and so is B_n^C. I want to show $\mu(S) \geq \mu(S\setminus A) + \mu(S\cap A)$. If $\mu(S) = \infty$, there is nothing to prove. Assume $\mu(S) < \infty$. Then apply Parts 9.13 and 9.12 to the outer measure $\mu\lfloor S$ in 9.15 and let $n \to \infty$. Thus $B_n \uparrow A$, $B_n^C \downarrow A^C$ and this yields $\mu(S) = (\mu\lfloor S)(A) + (\mu\lfloor S)(A^C) = \mu(S\cap A) + \mu(S\setminus A)$.

Therefore $A \in \mathscr{S}$ and this proves Parts 9.11, 9.12, and 9.13.

It only remains to verify the assertion about completeness. Letting G and F be as described above, let $S \subseteq \Omega$. I need to verify $\mu(S) \geq \mu(S\cap G) + \mu(S\setminus G)$. However,

$$\begin{aligned} \mu(S\cap G) + \mu(S\setminus G) &\leq \mu(S\cap F) + \mu(S\setminus F) + \mu(F\setminus G) \\ &= \mu(S\cap F) + \mu(S\setminus F) = \mu(S) \end{aligned}$$

because by assumption, $\mu(F\setminus G) \leq \mu(F) = 0$. ■

Corollary 9.8.5 *Completeness is the same as saying that if* $(E\setminus E')\cup(E'\setminus E) \subseteq N \in \mathscr{F}$ *and* $\mu(N) = 0$, *then if* $E \in \mathscr{F}$, *it follows that* $E' \in \mathscr{F}$ *also.*

Proof: If the new condition holds, then suppose $G \subseteq F$ where $\mu(F) = 0, F \in \mathscr{F}$. Then $\overbrace{(G\setminus F)}^{=\emptyset} \cup (F\setminus G) \subseteq F$ and $\mu(F)$ is given to equal 0. Therefore, $G \in \mathscr{F}$.

Now suppose the earlier version of completeness and let

$$(E\setminus E')\cup(E'\setminus E) \subseteq N \in \mathscr{F}$$

where $\mu(N) = 0$ and $E \in \mathscr{F}$. Then we know $(E\setminus E'), (E'\setminus E) \in \mathscr{F}$ and all have measure zero. It follows $E\setminus(E\setminus E') = E\cap E' \in \mathscr{F}$. Hence

$$E' = (E\cap E')\cup(E'\setminus E) \in \mathscr{F}$$ ■

The measure μ which results from the outer measure of Theorem 9.7.4 is called the Lebesgue Stieltjes measure associated with the integrator function F. Its properties will be discussed more in the next section.

9.9 Measurable Sets Include Borel Sets?

If you have an outer measure, it determines a measure. This section gives a very convenient criterion which allows you to conclude right away that the measure is a Borel measure.

Theorem 9.9.1 *Let* μ *be an outer measure on the subsets of* (X,d), *a metric space. If* $\mu(A\cup B) = \mu(A) + \mu(B)$ *whenever* $\operatorname{dist}(A,B) > 0$, *then the* σ *algebra of measurable sets* $\mathscr{S}$ *contains the Borel sets.*

Proof: It suffices to show that closed sets are in $\mathscr{S}$, the σ-algebra of measurable sets, because then the open sets are also in $\mathscr{S}$ and consequently $\mathscr{S}$ contains the Borel sets. Let K be closed and let S be a subset of Ω. Is $\mu(S) \geq \mu(S\cap K) + \mu(S\setminus K)$? It suffices to assume $\mu(S) < \infty$. Let $K_n \equiv \left\{x : \operatorname{dist}(x,K) \leq \frac{1}{n}\right\}$. By Lemma 9.0.4 on Page 241, $x \to \operatorname{dist}(x,K)$ is continuous and so K_n is a closed set having K as a proper subset. That in K_n^C is at a positive distance from K. By the assumption of the theorem,

$$\mu(S) \geq \mu((S\cap K)\cup(S\setminus K_n)) = \mu(S\cap K) + \mu(S\setminus K_n) \tag{9.16}$$

Now

$$\mu(S\setminus K_n) \le \mu(S\setminus K) \le \mu(S\setminus K_n) + \mu((K_n\setminus K)\cap S). \tag{9.17}$$

If $\lim_{n\to\infty} \mu((K_n\setminus K)\cap S) = 0$ then the theorem will be proved because this limit along with 9.17 implies $\lim_{n\to\infty}\mu(S\setminus K_n) = \mu(S\setminus K)$ and then taking a limit in 9.16, $\mu(S) \ge \mu(S\cap K) + \mu(S\setminus K)$ as desired. Therefore, it suffices to establish this limit.

Since K is closed, a point, $x \notin K$ must be at a positive distance from K and so

$$K_n \setminus K = \cup_{k=n}^{\infty} K_k \setminus K_{k+1}.$$

Therefore

$$\mu(S\cap(K_n\setminus K)) \le \sum_{k=n}^{\infty} \mu(S\cap(K_k\setminus K_{k+1})). \tag{9.18}$$

If

$$\sum_{k=1}^{\infty} \mu(S\cap(K_k\setminus K_{k+1})) < \infty, \tag{9.19}$$

then $\mu(S\cap(K_n\setminus K)) \to 0$ because it is dominated by the tail of a convergent series so it suffices to show 9.19.

$$\sum_{k=1}^{M} \mu(S\cap(K_k\setminus K_{k+1})) =$$

$$\sum_{k\ even,\, k\le M} \mu(S\cap(K_k\setminus K_{k+1})) + \sum_{k\ odd,\, k\le M} \mu(S\cap(K_k\setminus K_{k+1})). \tag{9.20}$$

By the construction, the distance between any pair of sets, $S\cap(K_k\setminus K_{k+1})$ for different even values of k is positive and the distance between any pair of sets, $S\cap(K_k\setminus K_{k+1})$ for different odd values of k is positive. Therefore,

$$\sum_{k\ even,\, k\le M} \mu(S\cap(K_k\setminus K_{k+1})) + \sum_{k\ odd,\, k\le M} \mu(S\cap(K_k\setminus K_{k+1})) \le$$

$$\mu\left(\bigcup_{k\ even,\, k\le M} (S\cap(K_k\setminus K_{k+1}))\right) + \mu\left(\bigcup_{k\ odd,\, k\le M} (S\cap(K_k\setminus K_{k+1}))\right)$$

$$\le \mu(S) + \mu(S) = 2\mu(S)$$

and so for all M, $\sum_{k=1}^{M}\mu(S\cap(K_k\setminus K_{k+1})) \le 2\mu(S)$ showing 9.19. ■

9.10 One Dimensional Lebesgue Stieltjes Measure

Now with these major results about measures, it is time to specialize to the outer measure of Theorem 9.7.4. The next theorem gives Lebesgue Stieltjes measure on $\mathbb{R}$. The conditions 9.21 and 9.22 given below are known respectively as inner and outer regularity.

Theorem 9.10.1 *Let $\mathcal{F}$ denote the σ algebra of Theorem 9.8.4, associated with the outer measure μ in Theorem 9.7.4, on which μ is a measure. Then every open interval is in $\mathcal{F}$. So are all open and closed sets and consequently all Borel sets. Furthermore, if E is any set in $\mathcal{F}$*

$$\mu(E) = \sup\{\mu(K) : K \text{ compact}, K\subseteq E\} \tag{9.21}$$

$$\mu(E) = \inf\{\mu(V) : V \text{ is an open set } V\supseteq E\} \tag{9.22}$$

If $E \in \mathscr{F}$, there exists F a countable union of compact sets, an F_σ set and a set G a countable intersection of open sets, a G_δ set such that $F \subseteq E \subseteq G$ but $\mu(G \setminus F) = 0$. Also μ is finite on compact sets.

Proof: By Theorem 9.7.4 and Theorem 9.9.1 the σ algebra includes the Borel sets $\mathscr{B}(\mathbb{R})$. However, note that $\mathscr{F}$ is complete and there is no such requirement for this measure on $\mathscr{B}(\mathbb{R})$. Thus it is reasonable to think that $\mathscr{F}$ could be larger than $\mathscr{B}(\mathbb{R})$.

Now consider the last claim about regularity. The assertion of outer regularity on $\mathscr{F}$ is not hard to get. Letting E be any set $\mu(E) < \infty$, there exist open intervals covering E denoted by $\{(a_i, b_i)\}_{i=1}^{\infty}$ such that

$$\mu(E) + \varepsilon > \sum_{i=1}^{\infty} F(b_i-) - F(a_i+) = \sum_{i=1}^{\infty} \mu(a_i, b_i) \geq \mu(V)$$

where V is the union of the open intervals just mentioned. Thus

$$\mu(E) \leq \mu(V) \leq \mu(E) + \varepsilon.$$

This shows outer regularity. If $\mu(E) = \infty$, there is nothing to show. Since μ is finite on intervals, it is σ finite. It follows from Theorem 9.4.6 that μ is inner regular also and the claim about approximation with F_σ and G_δ sets follows. ■

Definition 9.10.2 *When the integrator function is $F(x) = x$, the Lebesgue Stieltjes measure just discussed is known as one dimensional Lebesgue measure and is denoted as m.*

Proposition 9.10.3 *For m Lebesgue measure, $m([a,b]) = m((a,b)) = b - a$. Also m is translation invariant in the sense that if E is any Lebesgue measurable set, then $m(x+E) = m(E)$.*

Proof: The formula for the measure of an interval comes right away from Theorem 9.7.4. From this, it follows right away that whenever E is an interval, $m(x+E) = m(E)$. Every open set is the countable disjoint union of open intervals, so if E is an open set, then $m(x+E) = m(E)$. What about closed sets? First suppose H is a closed and bounded set. Then letting $(-n,n) \supseteq H$,

$$\mu(((-n,n) \setminus H) + x) + \mu(H + x) = \mu((-n,n) + x)$$

Hence, from what was just shown about open sets,

$$\begin{aligned} \mu(H) &= \mu((-n,n)) - \mu((-n,n) \setminus H) \\ &= \mu((-n,n) + x) - \mu(((-n,n) \setminus H) + x) = \mu(H + x) \end{aligned}$$

Therefore, the translation invariance holds for closed and bounded sets. If H is an arbitrary closed set, then

$$\mu(H + x) = \lim_{n \to \infty} \mu(H \cap [-n,n] + x) = \lim_{n \to \infty} \mu(H \cap [-n,n]) = \mu(H).$$

It follows right away that μ is translation invariant on F_σ and G_δ sets. Now using Theorem 9.10.1, if E is an arbitrary measurable set, there exist an F_σ set F and a G_δ set G such that $F \subseteq E \subseteq G$ and $m(F) = m(G) = m(E)$. Then

$$m(F) = m(x+F) \leq m(x+E) \leq m(x+G) = m(G) = m(E) = m(F). \blacksquare$$

9.11 Completion of a Measure Space

Next is the notion of the completion of a measure space. The idea is that you might not have completeness in your measure space but you can always complete it.

Definition 9.11.1 *Recall that a measure space $(\Omega,\mathcal{F},\lambda)$ is σ finite if there is a countable set $\{\Omega_n\}_{n=1}^{\infty}$ such that $\cup_n\Omega_n=\Omega$ and $\lambda(\Omega_n)<\infty$.*

The next theorem is like some earlier ones related to regularity including the approximation with G_δ and F_σ sets. The arguments are similar.

Theorem 9.11.2 *Let $(\Omega,\mathcal{F},\mu)$ be a measure space. Then there exists a measure space, $(\Omega,\mathcal{G},\lambda)$ satisfying*

1. *$(\Omega,\mathcal{G},\lambda)$ is a complete measure space.*
2. *$\lambda=\mu$ on $\mathcal{F}$*
3. *$\mathcal{G}\supseteq\mathcal{F}$*
4. *For every $E\in\mathcal{G}$ there exists $G\in\mathcal{F}$ such that $G\supseteq E$ and $\mu(G)=\lambda(E)$.*

   In addition to this, if $(\Omega,\mathcal{F},\mu)$ is σ finite, then the following approximation result holds.
5. *For every $E\in\mathcal{G}$ there exists $F\in\mathcal{F}$ and $G\in\mathcal{F}$ such that $F\subseteq E\subseteq G$ and*

$$\mu(G\setminus F)=\lambda(G\setminus F)=0 \tag{9.23}$$

   There is a unique complete measure space $(\Omega,\mathcal{G},\lambda)$ extending $(\Omega,\mathcal{F},\mu)$ which satisfies 9.23. In particular, there are no new sets if the original measure space was already complete.

Proof: Define the outer measure

$$\lambda(A)\equiv\inf\{\mu(E):E\in\mathcal{F},E\supseteq A\},\ \lambda(\emptyset)\equiv 0.$$

Denote by $\mathcal{G}$ the σ algebra of λ measurable sets. Then $(\Omega,\mathcal{G},\lambda)$ is complete by the general Caratheodory procedure presented earlier.

I claim that $\lambda=\mu$ on $\mathcal{F}$. If $A\in\mathcal{F}$,

$$\mu(A)\leq\inf\{\mu(E):E\in\mathcal{F},E\supseteq A\}\equiv\lambda(A)\leq\mu(A)$$

because $A\supseteq A$. Thus, these are all equal in the above and $\lambda=\mu$ on $\mathcal{F}$.

Why is $\mathcal{F}\subseteq\mathcal{G}$? Letting $\lambda(S)<\infty$, (There is nothing to prove if $\lambda(S)=\infty$.) let $G\in\mathcal{F}$ be such that $G\supseteq S$ and $\lambda(S)=\mu(G)$. This is possible because

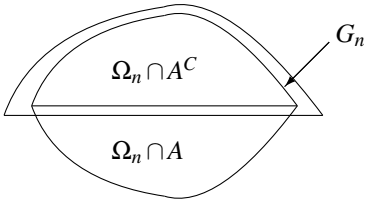
$$\lambda(S)\equiv\inf\{\mu(E):E\supseteq S\text{ and }E\in\mathcal{F}\}.$$

Then if $A\in\mathcal{F}$,

$$\begin{aligned}\lambda(S) &\leq \lambda(S\cap A)+\lambda\left(S\cap A^C\right)\leq\lambda(G\cap A)+\lambda\left(G\cap A^C\right)\\ &= \mu(G\cap A)+\mu\left(G\cap A^C\right)=\mu(G)=\lambda(S).\end{aligned}$$

Thus $\mathscr{F} \subseteq \mathscr{G}$.

Finally suppose μ is σ finite. Let $\Omega = \cup_{n=1}^{\infty} \Omega_n$ where the Ω_n are disjoint sets of $\mathscr{F}$ and $\mu(\Omega_n) < \infty$. If the Ω_n are not disjoint, replace Ω_n with $\Omega_n \setminus \cup_{k=1}^{n-1} \Omega_k$. Letting $A \in \mathscr{G}$, consider $A_n \equiv A \cap \Omega_n$. From what was just shown, there exists $G_n \supseteq A^C \cap \Omega_n$, $G_n \subseteq \Omega_n$ such that $\mu(G_n) = \lambda\left(A^C \cap \Omega_n\right), G_n \in \mathscr{F}$.



Since $\mu(\Omega_n) < \infty$, this implies

$$\lambda\left(G_n \setminus \left(A^C \cap \Omega_n\right)\right) = \lambda(G_n) - \lambda\left(A^C \cap \Omega_n\right) = 0.$$

Now $G_n^C \subseteq A \cup \Omega_n^C$ but $G_n \subseteq \Omega_n$ and so $G_n^C \subseteq A \cup \Omega_n$. Define $F_n \equiv G_n^C \cap \Omega_n \subseteq A_n$ and it follows $\lambda(A_n \setminus F_n) =$

$$\begin{aligned} \lambda\left(A \cap \Omega_n \setminus \left(G_n^C \cap \Omega_n\right)\right) &= \lambda(A \cap \Omega_n \cap G_n) = \lambda(A \cap G_n) = \lambda\left(G_n \setminus A^C\right) \\ &\leq \lambda\left(G_n \setminus \left(A^C \cap \Omega_n\right)\right) = 0. \end{aligned}$$

Letting $F = \cup_n F_n$, it follows that $F \in \mathscr{F}$ and

$$\lambda(A \setminus F) \leq \sum_{k=1}^{\infty} \lambda(A_k \setminus F_k) = 0.$$

Also, there exists $G_n \supseteq A_n$ such that $\mu(G_n) = \lambda(G_n) = \lambda(A_n)$. Since the measures are finite, it follows that $\lambda(G_n \setminus A_n) = 0$. Then letting $G = \cup_{n=1}^{\infty} G_n$, it follows that $G \supseteq A$ and

$$\begin{aligned} \lambda(G \setminus A) &= \lambda\left(\cup_{n=1}^{\infty} G_n \setminus \cup_{n=1}^{\infty} A_n\right) \\ &\leq \lambda\left(\cup_{n=1}^{\infty} (G_n \setminus A_n)\right) \leq \sum_{n=1}^{\infty} \lambda(G_n \setminus A_n) = 0. \end{aligned}$$

Thus $\mu(G \setminus F) = \lambda(G \setminus F) = \lambda(G \setminus A) + \lambda(A \setminus F) = 0$.

If you have $\left(\lambda', \mathscr{G}'\right)$ complete and satisfying 9.23, then letting $E \in \mathscr{G}'$, it follows from 5, that there exist $F, G \in \mathscr{F}$ such that

$$F \subseteq E \subseteq G,\ \mu(G \setminus F) = 0 = \lambda(G \setminus F).$$

Therefore, by completeness of the two measure spaces, $E \in \mathscr{G}$. The opposite inclusion is similar. Hence $\mathscr{G} = \mathscr{G}'$. If $E \in \mathscr{G}$, let $F \subseteq E \subseteq G$ where $\mu(G \setminus F) = 0$. Then

$$\lambda(E) \leq \mu(G) = \mu(F) = \lambda'(F) \leq \lambda'(E)$$

The opposite inequality holds by the same reasoning. Hence $\lambda = \lambda'$. If $(\Omega, \mathscr{F}, \mu)$ is already complete, then you could let this be $\left(\Omega, \mathscr{G}', \lambda'\right)$ and find that $\mathscr{G} = \mathscr{F} = \mathscr{G}'$. ■

Another useful result is the following.

Corollary 9.11.3 *Suppose, in the situation of Theorem 9.11.2, $f \geq 0$ and is $\mathscr{G}$ measurable. Then there exists $g, 0 \leq g \leq f$ and $f = g$ for all ω off a set of measure zero.*

Proof: Let $s_n \uparrow f$ where $s_n(\omega) = \sum_{i=1}^{m_n} c_i \mathcal{X}_{E_i}(\omega)$ for $\omega \in \Omega$. Then by the regularity assertion of this theorem, there exists $F_i \in \mathcal{F}$ such that $F_i \subseteq E_i$ and $\lambda(E_i \setminus F_i) = 0$. Then let $\hat{s}_n(\omega) = \sum_{i=1}^{m_n} c_i \mathcal{X}_{F_i}(\omega)$. Then $\hat{s}_n \leq s_n$ and letting $N = \cup_{n=1}^{\infty} \{\omega : s_n(\omega) \neq \hat{s}_n(\omega)\}$, it follows that $\lambda(N) = 0$ and for $\omega \notin N$,

$$\hat{s}_n(\omega) = s_n(\omega) \to f(\omega) = g(\omega).$$

Now let $g(\omega) \equiv \liminf_{n\to\infty} \hat{s}_n(\omega) \leq \lim_{n\to\infty} s(\omega) = f(\omega)$ and g is $\mathcal{F}$ measurable because if $g_n(\omega) = \inf\{\hat{s}_k : k \geq n\}$, this is $\mathcal{F}$ measurable since

$$g_n^{-1}((-\infty, a)) = \cup_{k\geq n} \hat{s}_k^{-1}(-\infty, a) \in \mathcal{F}$$

Now g being the limit of these g_n, it follows that g is also $\mathcal{F}$ measurable. ■

This will show that in most situations, you can simply modify your function on a set of measure zero and consider one which is $\mathcal{F}$ measurable.

Recall Corollary 9.4.9 about regularity. Then there is an easy corollary.

Corollary 9.11.4 *Let Ω be a complete metric space which is the countable union of compact sets K_n and suppose, for μ a Borel measure, $\mu(K_n)$ is finite. Then μ must be regular on $\mathcal{B}(\Omega)$. If $(\bar{\mu}, \mathcal{G})$ is the completion, then $\bar{\mu}$ is inner and outer regular on sets of $\mathcal{G}$. Also, if $E \in \mathcal{G}$, there are F_σ and G_δ sets, F, G respectively such that $\bar{\mu}(G \setminus F) = 0$ and $F \subseteq E \subseteq G$.*

9.12 Vitali Coverings

There is another covering theorem which may also be referred to as the Besicovitch covering theorem. At first, the balls will be closed but this assumption will be removed. Assume the following: $(X, \bar{\mu})$ is a finite dimensional normed linear space of dimension p and $\bar{\mu}$ is an outer measure on $\mathcal{P}(X)$. We really have in mind that X is $\mathbb{R}^p$ with some norm. Assume the following:

1. Let μ the measure determined by $\bar{\mu}$ on the σ algebra $\mathcal{S}$ which contains the Borel sets.

2. Or let μ be a measure on $\mathcal{S}$ where $\mathcal{S}$ contains the Borel sets and $\bar{\mu}$ is the outer measure determined by μ as described in Proposition 9.5.2. Always assume the following:

3. $\mu(B(\boldsymbol{x}, r)) < \infty$.

4. If $E \in \mathcal{S}$, then

$$\begin{aligned} \mu(E) &= \sup\{\mu(K) : K \subseteq E \text{ and } K \text{ is compact}\} \\ \mu(E) &= \inf\{\mu(V) : V \supseteq E \text{ and } V \text{ is open}\} \end{aligned}$$

If this measure μ is also complete, then recall it is termed a Radon measure.

Note that $\mathcal{S}$ is given to contain all closed sets and open sets. The above situation is very common. See Corollary 9.4.9 which gives 4 follows from 3. In fact, the above is the typical case for measures on finite dimensional spaces.

Definition 9.12.1 *A collection of balls, $\mathscr{F}$ covers a set E in the sense of Vitali if whenever $x \in E$ and $\varepsilon > 0$, there exists a ball $B \in \mathscr{F}$ whose center is x having diameter less than ε.*

I will give a proof of the following theorem.

Theorem 9.12.2 *Let E be a set with $\overline{\mu}(E) < \infty$ and either 1 or 2 along with the regularity conditions 3 and 4. Suppose $\mathscr{F}$ is a collection of closed balls which cover E in the sense of Vitali. Then there exists a sequence of disjoint balls $\{B_i\} \subseteq \mathscr{F}$ such that $\overline{\mu}\left(E \setminus \cup_{j=1}^{N} B_j\right) = 0$, $N \leq \infty$.*

Proof: Let N_p be the constant of the Besicovitch covering theorem, Theorem 4.5.8. Choose $r > 0$ such that $(1-r)^{-1}\left(1 - \frac{1}{2N_p+2}\right) \equiv \lambda < 1$. If $\overline{\mu}(E) = 0$, there is nothing to prove so assume $\overline{\mu}(E) > 0$. Let U_1 be an open set containing E with $(1-r)\mu(U_1) < \overline{\mu}(E)$ and $2\overline{\mu}(E) > \mu(U_1)$, and let $\mathscr{F}_1$ be those sets of $\mathscr{F}$ which are contained in U_1 whose centers are in E. Thus $\mathscr{F}_1$ is also a Vitali cover of E. Now by the Besicovitch covering theorem proved earlier, Theorem 4.5.8, there exist balls B, of $\mathscr{F}_1$ such that $E \subseteq \cup_{i=1}^{N_p} \{B : B \in \mathscr{G}_i\}$ where $\mathscr{G}_i$ consists of a collection of disjoint balls of $\mathscr{F}_1$. Therefore, $\overline{\mu}(E) \leq \sum_{i=1}^{N_p} \sum_{B \in \mathscr{G}_i} \mu(B)$ and so, for some $i \leq N_p$,

$$(N_p+1) \sum_{B \in \mathscr{G}_i} \mu(B) > \overline{\mu}(E).$$

It follows there exists a finite set of balls of $\mathscr{G}_i$, $\{B_1, \cdots, B_{m_1}\}$ such that

$$(N_p+1) \sum_{i=1}^{m_1} \mu(B_i) > \overline{\mu}(E) \tag{9.24}$$

and so

$$(2N_p+2) \sum_{i=1}^{m_1} \mu(B_i) > 2\overline{\mu}(E) > \mu(U_1).$$

Now 9.24 implies

$$\frac{\mu(U_1)}{2N_2+2} \leq \frac{2\overline{\mu}(E)}{2N_2+2} = \frac{\overline{\mu}(E)}{N_2+1} < \sum_{i=1}^{m_1} \mu(B_i).$$

Also U_1 was chosen such that $(1-r)\mu(U_1) < \overline{\mu}(E)$, and so

$$\lambda \overline{\mu}(E) \geq \lambda(1-r)\mu(U_1) = \left(1 - \frac{1}{2N_p+2}\right)\mu(U_1)$$

$$\geq \mu(U_1) - \sum_{i=1}^{m_1} \mu(B_i) = \mu(U_1) - \mu\left(\cup_{j=1}^{m_1} B_j\right)$$

$$= \mu\left(U_1 \setminus \cup_{j=1}^{m_1} B_j\right) \geq \overline{\mu}\left(E \setminus \cup_{j=1}^{m_1} B_j\right).$$

Since the balls are closed, you can consider the sets of $\mathscr{F}$ which have empty intersection with $\cup_{j=1}^{m_1} B_j$ and this new collection of sets will be a Vitali cover of $E \setminus \cup_{j=1}^{m_1} B_j$. Letting this collection of balls play the role of $\mathscr{F}$ in the above argument, and letting $E \setminus \cup_{j=1}^{m_1} B_j$ play

the role of E, repeat the above argument and obtain disjoint sets of $\mathscr{F}$, $\{B_{m_1+1},\cdots,B_{m_2}\}$, such that

$$\lambda\overline{\mu}\left(E\setminus\cup_{j=1}^{m_1}B_j\right)>\overline{\mu}\left(\left(E\setminus\cup_{j=1}^{m_1}B_j\right)\setminus\cup_{j=m_1+1}^{m_2}B_j\right)=\overline{\mu}\left(E\setminus\cup_{j=1}^{m_2}B_j\right),$$

and so $\lambda^2\overline{\mu}\ (E)>\overline{\mu}\ \left(E\setminus\cup_{j=1}^{m_2}B_j\right)$. Continuing in this way, yields a sequence of disjoint balls $\{B_i\}$ contained in $\mathscr{F}$ and $\overline{\mu}\ \left(E\setminus\cup_{j=1}^{N}B_j\right)\leq\overline{\mu}\ \left(E\setminus\cup_{j=1}^{m_k}B_j\right)<\lambda^k\overline{\mu}\ (E\ )$ for all k. If the process stops because E gets covered, then N is finite and if not, then $N=\infty$. Therefore, $\overline{\mu}\ \left(E\setminus\cup_{j=1}^{N}B_j\right)=0$ and this proves the Theorem. ■

It is not necessary to assume $\overline{\mu}(E)<\infty$. It is given that $\mu(B(\boldsymbol{x},R))<\infty$. Letting $C(\boldsymbol{x},r)$ be all $\boldsymbol{y}$ with $\|\boldsymbol{y}-\boldsymbol{x}\|=r$. Then there are only finitely many $r<R$ such that $\mu(C(\boldsymbol{x},r))\geq\frac{1}{n}$. Hence there are only countably many $r<R$ such that $\mu(C(\boldsymbol{x},r))>0$.

Corollary 9.12.3 *Let E nonempty set and either 1 or 2 along with the regularity conditions 3 and 4. Suppose $\mathscr{F}$ is a collection of closed balls which cover E in the sense of Vitali. Then there exists a sequence of disjoint balls $\{B_i\}\subseteq\mathscr{F}$ such that*

$$\overline{\mu}\left(E\setminus\cup_{j=1}^{N}B_j\right)=0,N\leq\infty$$

Proof: By 3, μ is finite on compact sets. Recall these are closed and bounded. There are at most countably many numbers, $\{b_i\}_{i=1}^{\infty}$ such that $\mu(C(\mathbf{0},b_i))>0$. It follows that there exists an increasing sequence of positive numbers, $\{r_i\}_{i=1}^{\infty}$ such that $\lim_{i\to\infty}r_i=\infty$ and $\mu(C(\mathbf{0},r_i))=0$. Now let

$$\begin{aligned}D_1&\equiv\{\boldsymbol{x}:\|\boldsymbol{x}\|<r_1\},D_2\equiv\{\boldsymbol{x}:r_1<\|\boldsymbol{x}\|<r_2\},\\ \cdots,D_m&\equiv\{\boldsymbol{x}:r_{m-1}<\|\boldsymbol{x}\|<r_m\},\cdots.\end{aligned}$$

Let $\mathscr{F}_m$ denote those closed balls of $\mathscr{F}$ which are contained in D_m. Then letting E_m denote $E\cap D_m$, $\mathscr{F}_m$ is a Vitali cover of $E_m,\overline{\mu}(E_m)<\infty$, and so by Theorem 9.12.2, there exists a countable sequence of balls from $\mathscr{F}_m$ $\left\{B_j^m\right\}_{j=1}^{N}$, such that $\overline{\mu}\left(E_m\setminus\cup_{j=1}^{N}B_j^m\right)=0$. Then consider the countable collection of balls, $\left\{B_j^m\right\}_{j,m=1}^{\infty}$.

$$\begin{aligned}\overline{\mu}\left(E\setminus\cup_{m=1}^{\infty}\cup_{j=1}^{N}B_j^m\right)&\leq\overline{\mu}\left(\cup_{j=1}^{\infty}\partial B(\mathbf{0},r_i)\right)+\\ +\sum_{m=1}^{\infty}\overline{\mu}\left(E_m\setminus\cup_{j=1}^{N}B_j^m\right)&=0\ ,N\leq\infty.\ \blacksquare\end{aligned}$$

If some E_m is empty, you could let your balls be the empty set.

You don't need to assume the balls are closed. In fact, the balls can be open, closed or anything in between and the same conclusion can be drawn provided you change the definition of a Vitali cover a little. For each point of the set covered, the covering includes all balls centered at that point having radius sufficiently small. In case that $\mu(C(\boldsymbol{x},r))=0$ for all $\boldsymbol{x},r$ where $C(\boldsymbol{x},r)\equiv\{\boldsymbol{y}:\|\boldsymbol{y}-\boldsymbol{x}\|=r\}$, no modification is necessary. This includes the case of Lebesgue measure. However, in the general case, consider the following modification of the notion of a Vitali cover.

Definition 9.12.4 *Suppose $\mathscr{F}$ is a collection of balls which cover E in the sense that for all $\varepsilon>0$ there are uncountably many balls of $\mathscr{F}$ centered at $\boldsymbol{x}$ having radius less than ε.*

Corollary 9.12.5 *Let 1 or 2 along with the regularity conditions 3 and 4. Suppose $\mathscr{F}$ is a collection of balls which cover E in the sense of Definition 9.12.4. Then there exists a sequence of disjoint balls, $\{B_i\} \subseteq \mathscr{F}$ such that $\overline{\mu}\left(E \setminus \cup_{j=1}^{N} B_j\right) = 0$ for $N \leq \infty$.*

Proof: Let $\boldsymbol{x} \in E$. Thus $\boldsymbol{x}$ is the center of arbitrarily small balls from $\mathscr{F}$. Since μ is finite on compact sets, only countably many can fail to have $\mu(\partial B(\boldsymbol{x}, r)) = 0$. Leave the balls out which have $\mu(\partial B(\boldsymbol{x}, r)) > 0$. Let $\mathscr{F}'$ denote the closures of the balls of $\mathscr{F}'$. Thus, for these balls, $\mu(\partial B(\boldsymbol{x}, r)) = 0$. Since for each $\boldsymbol{x} \in E$ there are only countably many exceptions, $\mathscr{F}'$ is still a Vitali cover of E. Therefore, by Corollary 9.12.3 there is a disjoint sequence of these balls of $\mathscr{F}'$, $\left\{\overline{B_i}\right\}_{i=1}^{\infty}$ for which $\overline{\mu}\left(E \setminus \cup_{j=1}^{N} \overline{B_j}\right) = 0$. However, since their boundaries have μ measure zero, it follows $\overline{\mu}\left(E \setminus \cup_{j=1}^{N} B_j\right) = 0$, $N \leq \infty$. ■

9.13 Differentiation of Increasing Functions

As a spectacular application of the covering theorem, is the famous theorem that an increasing function has a derivative a.e. Here the a.e. refers to Lebesgue measure, the Stieltjes measure from the increasing function $F(x) = x$.

Definition 9.13.1 *The Dini derivates are as follows. In these formulas, f is a real valued function defined on $\mathbb{R}$.*

$$\begin{aligned}
D^+ f(x) &\equiv \lim_{r \to 0+} \left(\sup_{0<u\leq r} \frac{f(x+u) - f(x)}{u} \right) \equiv \lim \sup_{u \to 0+} \frac{f(x+u) - f(x)}{u}, \\
D_+ f(x) &\equiv \lim_{r \to 0+} \left(\inf_{0<u\leq r} \frac{f(x+u) - f(x)}{u} \right) \equiv \lim \inf_{u \to 0+} \frac{f(x+u) - f(x)}{u}, \\
D^- f(x) &\equiv \lim_{r \to 0+} \left(\sup_{0<u\leq r} \frac{f(x) - f(x-u)}{u} \right) \equiv \lim \sup_{u \to 0+} \frac{f(x) - f(x-u)}{u}, \\
D_- f(x) &\equiv \lim_{r \to 0+} \left(\inf_{0<u\leq r} \frac{f(x) - f(x-u)}{u} \right) \equiv \lim \inf_{u \to 0+} \frac{f(x) - f(x-u)}{u}.
\end{aligned}$$

Lemma 9.13.2 *The function $f : \mathbb{R} \to \mathbb{R}$ has a derivative if and only if all the Dini derivates are equal.*

Proof: If $D^+ f(x) = D_+ f(x)$, then if u is small enough, let y_n be a decreasing sequence converging to x. Then

$$0 = D^+ f(x) - D_+ f(x) \geq \lim \sup_{n \to \infty} \frac{f(y_n) - f(x)}{y_n - x} - \lim \inf_{n \to \infty} \frac{f(y_n) - f(x)}{y_n - x}$$

and so the limit of the difference quotient exists for any such $\{y_n\}$. Thus the derivative from the right exists at x. Therefore, $D^+ f(x) > D_+ f(x)$ if and only if there is no right derivative. Similarly $D^- f(x) > D_- f(x)$ if and only if there is no derivative from the left at x. Also, there is a derivative if and only if there is a derivative from the left, right and the two are equal. This happens when $D^+ f(x) = D_- f(x) = D^- f(x) = D_+ f(x)$. Thus this happens if and only if all Dini derivates are equal. ■

The Lebesgue measure of single points is 0 and so we do not need to worry about whether the intervals are closed in using Corollary 9.12.3.

Let $\Delta f(I) = f(b) - f(x)$ or $f(x) - f(a)$ if I is an interval having end points $a < b$ with x the midpoint. Now suppose $\{J_j\}$ are disjoint intervals contained in I. Then, since f is increasing, $\Delta f(I) \geq \sum_j \Delta f(J_j)$. In this notation, the above lemma implies that if $D^- f(x) > b$ or $D^+ f(x) > b$, then for each $\varepsilon > 0$ there is an interval J of length less than ε which is centered at x and $\frac{\Delta f(J)}{(1/2)m(J)} > b$ where $m(J)$ is the Lebesgue measure of J which is the length of J. If either $D_- f(x)$ or $D_+ f(x) < a$, the above lemma implies that for each $\varepsilon > 0$ there exists I centered at x with $|I| < \varepsilon$ and $\frac{\Delta f(I)}{(1/2)m(I)} < a$. For example, if $D^- f(x) < a$, there exists a sequence $y_n \uparrow x$ with

$$\frac{f(y_n) - f(x)}{y_n - x} = \frac{f(x) - f(y_n)}{x - y_n} < a$$

so let I_n be the interval centered at x which has left end point y_n.

Note that the set of jumps J of an increasing function must be countable because these jumps determine disjoint open intervals of the form $(f(x-), f(x+))$ for $x \in J$ and each must contain a rational number of which, there are only countably many.

Lemma 9.13.3 *An increasing function f is Borel measurable and its derivates are Borel measurable functions.*

Proof: The set of jumps J is countable so it is a Borel set of measure zero, an F_σ set. Since f is increasing, its only points of discontinuity are points where it has a jump. Hence it is continuous off this set J. $f^{-1}([c,\infty))$ is an interval of the form $[d,\infty)$ or (d,∞). Thus f is Borel measurable. Consider $D^+ f$ for $x \notin J$. Let $g_r(x) \equiv \sup_{0<u\leq r} \frac{f(x+u)-f(x)}{u}$. Thus it is the supremum of functions continuous on J^C. Hence if $x \in g_r^{-1}(c,\infty), c \geq 0$, and $x \notin J^C$, $\frac{f(x+u)-f(x)}{u} > c$ for some $0 < u \leq r$. It follows that, since f is continuous at x, if $\hat{x}$ is close enough to x, it is also true that $\frac{f(\hat{x}+u)-f(\hat{x})}{u} > c$. Thus if $x \in g_r^{-1}(c,\infty) \cap J^C$, then for some δ_x small enough, $(x-\delta_x, x+\delta_x) \subseteq g_r^{-1}(c,\infty)$. Hence, $g_r^{-1}(c,\infty) \cap J^C$ is the intersection of an open set, the union of the intervals $(x-\delta_x, x+\delta_x)$ for $x \in g_r^{-1}(c,\infty) \cap J^C$, with J^C a Borel set. It follows that g_r is decreasing in r and is measurable because, since J is countable, $[g_r > c]$ is the union of a countable set with a Borel set. Thus for all x,

$$D^+ f(x) = \lim_{r\to 0+} \left(\sup_{0<u\leq r} \frac{f(x+u)-f(x)}{u} \right) = \lim_{r\to 0+} (g_r(x)) = \lim_{r_n \to 0} g_{r_n}(x)$$

where r_n is a decreasing sequence converging to 0. It follows that $x \to D^+ f(x)$ is Borel measurable as claimed because it is the limit of Borel measurable functions. Similar reasoning shows that the other derivates are measurable also. ■

Theorem 9.13.4 *Let $f : \mathbb{R} \to \mathbb{R}$ be increasing. Then $f'(x)$ exists for all x off a set of measure zero.*

Proof: Let N_{ab} for $0 < a < b$ denote either

$$\{x : D^+ f(x) > b > a > D_+ f(x)\}, \{x : D^- f(x) > b > a > D_- f(x)\},$$

or

$$\{x : D^- f(x) > b > a > D_+ f(x)\}, \{x : D^+ f(x) > b > a > D_- f(x)\}$$

From the above lemma, N_{ab} is measurable. Assume that N_{ab} is bounded and let V be open with $V \supseteq N_{ab}$, $m(N_{ab}) + \varepsilon > m(V)$. By Corollary 9.12.3 and the above discussion, there are open, disjoint intervals $\{I_n\}$, each centered at a point of N_{ab} such that

$$\frac{2\Delta f(I_n)}{m(I_n)} < a,\ m(N_{ab}) = m(N_{ab} \cap \cup_i I_i) = \sum_i m(N_{ab} \cap I_i)$$

Now do for $N_{ab} \cap I_i$ what was just done for N_{ab} and get disjoint intervals J_i^j contained in I_i with

$$\frac{2\Delta f\left(J_i^j\right)}{m\left(J_i^j\right)} > b,\ m(N_{ab} \cap I_i) = \sum_j m\left(N_{ab} \cap I_i \cap J_i^j\right)$$

Then

$$a(m(N_{ab}) + \varepsilon) > am(V) \geq a\sum_i m(I_i) > \sum_i 2\Delta f(I_i) \geq \sum_i \sum_j 2\Delta f\left(J_i^j\right)$$

$$\geq b\sum_i \sum_j m\left(J_i^j\right) \geq b\sum_i \sum_j m\left(J_i^j \cap N_{ab}\right) = b\sum_i m(N_{ab} \cap I_i) = bm(N_{ab})$$

Since ε is arbitrary and $a < b$, this shows $m(N_{ab}) = 0$. If N_{ab} is not bounded, apply the above to $N_{ab} \cap (-r, r)$ and conclude this has measure 0. Hence so does N_{ab}.

The countable union of N_{ab} for a, b positive rational and N_{ab} defined in any of the above ways is an exceptional set off which $D^+ f(x) = D_+ f(x) \geq D^- f(x) \geq D_- f(x) \geq D^+ f(x)$ and so these are all equal. This shows that off a set of measure zero, the function has a derivative a.e. ■

9.14 Exercises

1. Suppose you have $(X, \mathscr{F}, \mu)$ where $\mathscr{F} \supseteq \mathscr{B}(X)$ and also $\mu(B(x_0, r)) < \infty$ for all $r > 0$. Let $S(x_0, r) \equiv \{x \in X : d(x, x_0) = r\}$. Show that

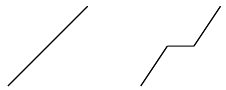
$$\{r > 0 : \mu(S(x_0, r)) > 0\}$$

cannot be uncountable. Explain why there exists a strictly increasing sequence $r_n \to \infty$ such that $\mu(x : d(x, x_0) = r_n) = 0$. In other words, the skin of the ball has measure zero except for possibly countably many values of the radius r.

2. Lebesgue measure was discussed. Recall that $m((a,b)) = b - a$ and it is defined on a σ algebra which contains the Borel sets, more generally on $\mathscr{P}(\mathbb{R})$. Also recall that m is translation invariant. Let $x \sim y$ if and only if $x - y \in \mathbb{Q}$. Show this is an equivalence relation. Now let W be a set of positive measure which is contained in $(0,1)$. For $x \in W$, let $[x]$ denote those $y \in W$ such that $x \sim y$. Thus the equivalence classes partition W. Use axiom of choice to obtain a set $S \subseteq W$ such that S consists of exactly one element from each equivalence class. Let $\mathbb{T}$ denote the rational numbers in $[-1, 1]$. Consider $\mathbb{T} + S \subseteq [-1, 2]$. Explain why $\mathbb{T} + S \supseteq W$. For $\mathbb{T} \equiv \{r_j\}$, explain why the sets $\{r_j + S\}_j$ are disjoint. Now suppose S is measurable. Then show that you have a contradiction if $m(S) = 0$ since $m(W) > 0$ and you also have a contradiction if $m(S) > 0$ because $\mathbb{T} + S$ consists of countably many disjoint sets. Explain why S cannot be measurable. Thus there exists $T \subseteq \mathbb{R}$ such that $m(T) < m(T \cap S) + m\left(T \cap S^C\right)$. Is there an open interval (a,b) such that if $T = (a,b)$, then the above inequality holds?

3. Consider the following nested sequence of compact sets, $\{P_n\}$.Let $P_1 = [0,1]$, $P_2 = \left[0,\frac{1}{3}\right] \cup \left[\frac{2}{3},1\right]$, etc. To go from P_n to P_{n+1}, delete the open interval which is the middle third of each closed interval in P_n. Let $P = \cap_{n=1}^{\infty} P_n$. By the finite intersection property of compact sets, $P \neq \emptyset$. Show $m(P) = 0$. If you feel ambitious also show there is a one to one onto mapping of $[0,1]$ to P. The set P is called the Cantor set. Thus, although P has measure zero, it has the same number of points in it as $[0,1]$ in the sense that there is a one to one and onto mapping from one to the other. **Hint:** There are various ways of doing this last part but the most enlightenment is obtained by exploiting the topological properties of the Cantor set rather than some silly representation in terms of sums of powers of two and three. All you need to do is use the Schroder Bernstein theorem and show there is an onto map from the Cantor set to $[0,1]$. If you do this right and remember the theorems about characterizations of compact metric spaces, Proposition 3.5.8 on Page 78, you may get a pretty good idea why every compact metric space is the continuous image of the Cantor set.

4. Consider the sequence of functions defined in the following way. Let $f_1(x) = x$ on $[0,1]$. To get from f_n to f_{n+1}, let $f_{n+1} = f_n$ on all intervals where f_n is constant. If f_n is nonconstant on $[a,b]$, let $f_{n+1}(a) = f_n(a)$, $f_{n+1}(b) = f_n(b)$, f_{n+1} is piecewise linear and equal to $\frac{1}{2}(f_n(a) + f_n(b))$ on the middle third of $[a,b]$. Sketch a few of these and you will see the pattern. The process of modifying a nonconstant section of the graph of this function is illustrated in the following picture.



Show $\{f_n\}$ converges uniformly on $[0,1]$. If $f(x) = \lim_{n\to\infty} f_n(x)$, show that $f(0) = 0$, $f(1) = 1$, f is continuous, and $f'(x) = 0$ for all $x \notin P$ where P is the Cantor set of Problem 3. This function is called the Cantor function.It is a very important example to remember. Note it has derivative equal to zero a.e. and yet it succeeds in climbing from 0 to 1. Explain why this interesting function cannot be recovered by integrating its derivative. (It is not absolutely continuous, explained later.) **Hint:** This isn't too hard if you focus on getting a careful estimate on the difference between two successive functions in the list considering only a typical small interval in which the change takes place. The above picture should be helpful.

5. ↑ This problem gives a very interesting example found in the book by McShane [40]. Let $g(x) = x + f(x)$ where f is the strange function of Problem 4. Let P be the Cantor set of Problem 3. Let $[0,1] \setminus P = \cup_{j=1}^{\infty} I_j$ where I_j is open and $I_j \cap I_k = \emptyset$ if $j \neq k$. These intervals are the connected components of the complement of the Cantor set. Show $m(g(I_j)) = m(I_j)$ so $m(g(\cup_{j=1}^{\infty} I_j)) = \sum_{j=1}^{\infty} m(g(I_j)) = \sum_{j=1}^{\infty} m(I_j) = 1$. Thus $m(g(P)) = 1$ because $g([0,1]) = [0,2]$. By Problem 2 there exists a set, $A \subseteq g(P)$ which is non measurable. Define $\phi(x) = \mathcal{X}_A(g(x))$. Thus $\phi(x) = 0$ unless $x \in P$. Tell why ϕ is measurable. (Recall $m(P) = 0$ and Lebesgue measure is complete.) Now show that $\mathcal{X}_A(y) = \phi(g^{-1}(y))$ for $y \in [0,2]$. Tell why g is strictly increasing and g^{-1} is continuous but $\phi \circ g^{-1}$ is not measurable. (This is an example of measurable $\circ$ continuous $\neq$ measurable.) Show there exist Lebesgue measurable sets which are not Borel measurable. **Hint**: The function, ϕ is Lebesgue measurable. Now recall that Borel $\circ$ measurable = measurable.

6. Show that every countable set of real numbers is of Lebesgue measure zero.

7. Review the Cantor set in Problem 12 on Page 178. You deleted middle third open intervals. Show that you can take out open intervals in the middle which are not necessarily middle thirds, and end up with a set C which has Lebesgue measure equal to $1-\varepsilon$. Also show if you can that there exists a continuous and one to one map $f : C \to J$ where J is the usual Cantor set of Problem 12 which also has measure 0.

8. Recall that every bounded variation function is the difference of two increasing functions. Show that every bounded variation function has a derivative a.e. For a discussion of these, see Definition 11.8.1 on Page 333 below if you have not seen it already.

9. Suppose you have a π system $\mathscr{K}$ of sets of Ω and suppose $\mathscr{G} \supseteq \mathscr{K}$ and that $\mathscr{G}$ is closed with respect to complements and that whenever $\{F_k\}$ is a decreasing sequence of sets of $\mathscr{G}$ it follows that $\cap_k F_k \in \mathscr{G}$. Show that then $\mathscr{G}$ contains $\sigma(\mathscr{K})$. This is an alternative formulation of Dynkin's lemma. It was shown after the Dynkin lemma that closure with respect to countable intersections is equivalent.

10. For $\boldsymbol{x} \in \mathbb{R}^p$ to be in $\prod_{i=1}^p A_i$, it means that the i^{th} component of $\boldsymbol{x}$, x_i is in A_i for each i. Now for $\prod_{i=1}^p (a_i, b_i) \equiv R$, let $V(R) = \prod_{i=1}^p (b_i - a_i)$. Next, for $A \in \mathscr{P}(\mathbb{R}^p)$ let

$$\mu(A) \equiv \inf\left\{\sum_k V\left(R^k\right) : A \subseteq \cup_k R^k\right\}$$

This is just like one dimensional Lebesgue measure except that instead of open intervals, we are using open boxes R^k. Show the following.

(a) μ is an outer measure.

(b) $\mu\left(\prod_{i=1}^p [a_i, b_i]\right) = \prod_{i=1}^p (b_i - a_i) = \mu\left(\prod_{i=1}^p (a_i, b_i)\right)$.

(c) If $\operatorname{dist}(A, B) > 0$, then $\mu(A) + \mu(B) = \mu(A \cup B)$ so $\mathscr{B}(\mathbb{R}^p) \subseteq \mathscr{F}$ the set of sets measurable with respect to this outer measure μ.

This is Lebesgue measure on $\mathbb{R}^p$. **Hint:** Suppose for some j, $b_j - a_j < \varepsilon$. Show that $\mu\left(\prod_{i=1}^p (a_i, b_i)\right) \leq \varepsilon \prod_{i \neq j} (b_i - a_i)$. Now use this to show that if you have a covering by finitely many open boxes, such that the sum of their volumes is less than some number, you can replace with a covering of open boxes which also has the sum of their volumes less than that number but which has each box with sides less than δ. To do this, you might consider replacing each box in the covering with 2^{mp} open boxes obtained by bisecting each side m times where m is small enough that each little box has sides smaller than $\delta/2$ in each of the finitely many boxes in the cover and then fatten each of these just a little to cover up what got left out and retain the sum of the volumes of the little boxes to still be less than the number you had.

11. ↑Show that Lebesgue measure defined in the above problem is both inner and outer regular and is translation invariant.

12. Let $(\Omega, \mathscr{F}, \mu)$ be a measure space and let $s(\omega) = \sum_{i=0}^n c_i \mathscr{X}_{E_i}(\omega)$ where the E_i are distinct measurable sets but the c_i might not be. Thus the c_i are the finitely many values of s. Say each $c_i \geq 0$ and $c_0 = 0$. Define $\int s d\mu$ as $\sum_i c_i \mu(E_i)$. Show that this is

well defined and that if you have $s(\omega)=\sum_{i=1}^{n}c_i\mathcal{X}_{E_i}(\omega),t(\omega)=\sum_{j=1}^{m}d_j\mathcal{X}_{F_j}(\omega)$, then for a,b nonnegative numbers, $as(\omega)+bt(\omega)$ can be written also in this form and that $\int(as+bt)\,d\mu=a\int s\,d\mu+b\int t\,d\mu$. **Hint:** $s(\omega)=\sum_i\sum_j c_i\mathcal{X}_{E_i\cap F_j}(\omega)=\sum_j\sum_i c_i\mathcal{X}_{E_i\cap F_j}(\omega)$ and $(as+bt)(\omega)=\sum_j\sum_i(ac_i+bd_j)\,\mathcal{X}_{E_i\cap F_j}(\omega)$.

13. ↑Having defined the integral of nonnegative simple functions in the above problem, letting f be nonnegative and measurable. Define

$$\int f\,d\mu\equiv\sup\left\{\int s\,d\mu:0\le s\le f,s\text{ simple}\right\}.$$

Show that if f_n is nonnegative and measurable and $n\to f_n(\omega)$ is increasing, show that for $f(\omega)=\lim_{n\to\infty}f_n(\omega)$, it follows that $\int f\,d\mu=\lim_{n\to\infty}\int f_n\,d\mu$. **Hint:** Show $\int f_n\,d\mu$ is increasing to something $\alpha\le\infty$. Explain why $\int f\,d\mu\ge\alpha$. Now pick a nonnegative simple function $s\le f$. For $r\in(0,1),[f_n>rs]\equiv E_n$ is increasing in n and $\cup_n E_n=\Omega$. Tell why $\int f_n\,d\mu\ge\int\mathcal{X}_{E_n}f_n\,d\mu\ge r\int s\,d\mu$. Let $n\to\infty$ and show that $\alpha\ge r\int s\,d\mu$. Now explain why $\alpha\ge r\int f\,d\mu$. Since r is arbitrary, $\alpha\ge\int f\,d\mu\ge\alpha$.

14. ↑Show that if f,g are nonnegative and measurable and $a,b\ge0$, then

$$\int(af+bg)\,d\mu=a\int f\,d\mu+b\int g\,d\mu$$

9.15 Multifunctions and Their Measurability

This is an introduction to the idea of measurable multifunctions. This is a very important topic which has surprising usefulness in nonlinear analysis and other areas and not enough attention is payed to it. As an application, I will give a proof of Kuratowski's theorem and also an interesting fixed point result in which the fixed point is a measurable function of ω in a measure space. One of the main references for this material is the book Papageorgiu and Hu [31] where you can find more of this kind of thing.

9.15.1 The General Case

Let X be a separable complete metric space and let $(\Omega,\mathcal{F})$ be a set and a σ algebra of subsets of Ω. A multifunction, is a map from Ω to the nonempty subsets of X. Thus Γ is a multifunction if for each ω, $\Gamma(\omega)\ne\emptyset$. For more on the theorems presented in this section, see [31].

Definition 9.15.1 *Define* $\Gamma^{-}(S)\equiv\{\omega\in\Omega:\Gamma(\omega)\cap S\ne\emptyset\}$. *When*

$$\Gamma^{-}(U)\in\mathcal{F}$$

for all U open, we say that Γ is measurable.

More can be said than what follows, but the following is the essential idea for a measurable multifunction.

Theorem 9.15.2 *The following are equivalent for any measurable space consisting only of a set Ω and a σ algebra $\mathcal{F}$. Here nothing is known about $\Gamma(\omega)$ other than that is a nonempty set.*

1. *For all U open in $X, \Gamma^{-}(U) \in \mathscr{F}$ where $\Gamma^{-}(U) \equiv \{\omega : \Gamma(\omega) \cap U \neq \emptyset\}$*

2. *There exists a sequence, $\{\sigma_n\}$ of measurable functions satisfying $\sigma_n(\omega) \in \Gamma(\omega)$ such that for all $\omega \in \Omega, \overline{\Gamma(\omega)} = \overline{\{\sigma_n(\omega) : n \in \mathbb{N}\}}$. These functions are called measurable selections.*

Proof: First 1.) $\Rightarrow$ 2.). A measurable selection will be obtained in $\overline{\Gamma(\omega)}$. Let $D \equiv \{x_n\}_{n=1}^{\infty}$ be a countable dense subset of X. For $\omega \in \Omega$, let $\psi_1(\omega) = x_n$ where n is the smallest integer such that $\Gamma(\omega) \cap B(x_n, 1) \neq \emptyset$. Therefore, $\psi_1(\omega)$ has countably many values, $x_{n_1}, x_{n_2}, \cdots$ where $n_1 < n_2 < \cdots$. Now the set on which ψ_1 has the value x_n is as follows: $\{\omega : \psi_1 = x_n\} =$

$$\{\omega : \Gamma(\omega) \cap B(x_n, 1) \neq \emptyset\} \cap [\Omega \setminus \cup_{k<n} \{\omega : \Gamma(\omega) \cap B(x_k, 1) \neq \emptyset\}] \in \mathscr{F}.$$

Thus ψ_1 is measurable and $\operatorname{dist}(\psi_1(\omega), \Gamma(\omega)) < 1$. Let $\Omega_n \equiv \{\omega \in \Omega : \psi_1(\omega) = x_n\}$. Then $\Omega_n \in \mathscr{F}$ and $\Omega_n \cap \Omega_m = \emptyset$ for $n \neq m$ and $\cup_{n=1}^{\infty} \Omega_n = \Omega$ because if ω is given, $\Gamma(\omega)$ does intersect some $B(x_n, 1)$. Let

$$D_n \equiv \{x_k \in D : x_k \in B(x_n, 1)\}.$$

Now for each n, and $\omega \in \Omega_n$, let $\psi_2(\omega) = x_k$ where k is the smallest index such that $x_k \in D_n$ and $B\left(x_k, \frac{1}{2}\right) \cap \Gamma(\omega) \neq \emptyset$. Thus

$$\operatorname{dist}(\psi_2(\omega), \Gamma(\omega)) < \frac{1}{2}, \ d(\psi_2(\omega), \psi_1(\omega)) < 1. \tag{9.25}$$

This defines $\psi_2(\omega)$ on Ω_n and so it defines ψ_2 on Ω satisfying 9.25. Continue this way, obtaining ψ_k a measurable function such that

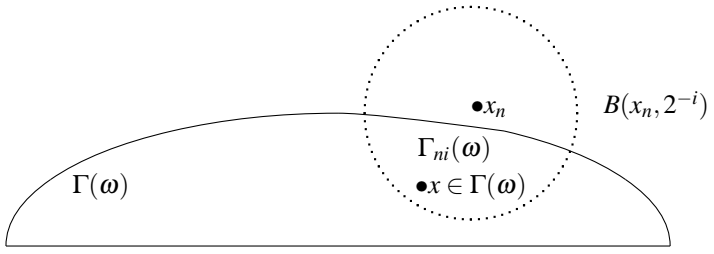
$$\operatorname{dist}(\psi_k(\omega), \Gamma(\omega)) < \frac{1}{2^{k-1}}, \ d\left(\psi_{k+1}(\omega), \psi_k(\omega)\right) < \frac{1}{2^{k-2}}.$$

Then for each $\omega, \{\psi_k(\omega)\}$ is a Cauchy sequence of measurable functions converging to a point, $\sigma(\omega) \in \overline{\Gamma(\omega)}$. This has shown that if Γ is measurable, there exists a measurable selection, $\sigma(\omega) \in \overline{\Gamma(\omega)}$. Of course, if $\Gamma(\omega)$ is closed, then $\sigma(\omega) \in \Gamma(\omega)$. Note that this had nothing to do with any measure.

It remains to show that there exists a sequence of these measurable selections σ_n such that the conclusion of 2.) holds. To do this define for a single $\omega \in \Omega$

$$\Gamma_{ni}(\omega) \equiv \begin{cases} \Gamma(\omega) \cap B\left(x_n, 2^{-i}\right) \text{ if } \Gamma(\omega) \cap B\left(x_n, 2^{-i}\right) \neq \emptyset \\ \Gamma(\omega) \text{ otherwise when there is empty intersection} \end{cases}.$$

The following picture illustrates $\Gamma_{ni}(\omega)$ when ω is such that there is nonempty intersection. Also, given $x \in \Gamma(\omega)$, and i, there is x_n from the countable dense set such that the situation of the picture occurs.



Is Γ_{ni} measurable? If so, then from the above, it has a measurable selection σ_{ni} and

the set of these σ_{ni} must have the property that $\{\sigma_{ni}(\omega)\}_{n,i}$ is dense in $\Gamma(\omega)$ for each ω.

Let U be open. Then

$$\begin{aligned}\{\omega : \Gamma_{ni}(\omega) \cap U \neq \emptyset\} &= \left\{\omega : \Gamma(\omega) \cap B\left(x_n, 2^{-i}\right) \cap U \neq \emptyset\right\} \cup \\ &\left[\left\{\omega : \Gamma(\omega) \cap B\left(x_n, 2^{-i}\right) = \emptyset\right\} \cap \{\omega : \Gamma(\omega) \cap U \neq \emptyset\}\right] \\ &= \left\{\omega : \Gamma(\omega) \cap B\left(x_n, 2^{-i}\right) \cap U \neq \emptyset\right\} \cup \\ &\left[\left(\Omega \setminus \left\{\omega : \Gamma(\omega) \cap B\left(x_n, 2^{-i}\right) \neq \emptyset\right\}\right) \cap \{\omega : \Gamma(\omega) \cap U \neq \emptyset\}\right],\end{aligned}$$

a measurable set. Thus Γ_{ni} is measurable as hoped.

By what was just shown, there exists σ_{ni}, a measurable function such that $\sigma_{ni}(\omega) \in \overline{\Gamma_{ni}(\omega)} \subseteq \overline{\Gamma(\omega)}$ for all $\omega \in \Omega$. If $x \in \overline{\Gamma(\omega)}$, then $x \in B\left(x_n, 2^{-(i+1)}\right)$ whenever x_n is close enough to x. Thus both $x, \sigma_{n(i+1)}(\omega)$ are in $\overline{B\left(x_n, 2^{-(i+1)}\right)}$ and so $\left|\sigma_{n(i+1)}(\omega) - x\right| < 2^{-i}$. It follows that condition 2.) holds with the countable dense subset of $\overline{\Gamma(\omega)}$ being the $\{\sigma_{ni}(\omega)\}$. Note that this had nothing to do with a measure.

Now consider why 2.)$\Rightarrow$ 1.). We have $\{\sigma_n(\omega)\} \subseteq \Gamma(\omega)$ and σ_n is measurable and $\overline{\cup_n \sigma_n(\omega)}$ equals $\overline{\Gamma(\omega)}$. Why is Γ a measurable multifunction? Let U be an open set

$$\begin{aligned}\Gamma^-(U) &\equiv \{\omega : \Gamma(\omega) \cap U \neq \emptyset\} = \left\{\omega : \overline{\Gamma(\omega)} \cap U \neq \emptyset\right\} \\ &= \cup_n \sigma_n^{-1}(U) \in \mathscr{F} \ \blacksquare\end{aligned}$$

For much more on multi-functions, you should see the book by Hu and Papageorgiou. [31] The above proof follows the presentation in this book but there is more to be seen there where complete measures are included in the theory and an equivalence is shown between strong measurability, about to be discussed, and measurability without an assumption that the multifunction has compact values.

9.15.2 A Special Case When $\Gamma(\omega)$ Compact

Measurability is a statement that $\Gamma^-(U) \in \mathscr{F}$ whenever U is open.

Definition 9.15.3 *A multifunction Γ is strongly measurable if $\Gamma^-(F) \in \mathscr{F}$ for all F closed.*

Observation 9.15.4 *If Γ is strongly measurable, then it is measurable because if you have U open in a metric space, it is the countable union of closed sets F_n. Hence $\Gamma^-(U) = \cup_k \Gamma^-(F_k) \in \mathscr{F}$.*

Now suppose $\Gamma(\omega)$ is compact for every ω and that $\Gamma^-(U) \in \mathscr{F}$ for every U open. Then let F be a closed set and let $\{U_n\}$ be a decreasing sequence of open sets whose intersection equals F such that also, for all n, $U_n \supseteq \overline{U_{n+1}}$. Then

$$\Gamma(\omega) \cap F = \cap_n \Gamma(\omega) \cap U_n = \cap_n \Gamma(\omega) \cap \overline{U_n}$$

Now because of compactness, the set on the left is nonempty if and only if each set $\Gamma(\omega) \cap \overline{U_n}$ on the right is also nonempty. Thus $\Gamma^-(F) = \cap_n \Gamma^-(U_n) \in \mathscr{F}$. It follows that in this special case, the two conditions, measurability and strong measurability are equivalent. Note that there is no condition on measures or completeness or any such thing. This proves the following proposition.

Proposition 9.15.5 *Let X be a Polish space and let $\Gamma : X \to \mathscr{P}(X)$ have compact values. Then Γ is measurable if and only if it is strongly measurable, the latter being the statement that $\Gamma^-(C)$ is measurable whenever C is closed.*

Recall how if a function f is measurable, then $f^{-1}(\text{Borel set}) \in \mathscr{F}$. Something like this happens in case Γ is strongly measurable. Let Γ be strongly measurable. Let $\mathscr{G}$ be the sets G such that $\Gamma^-(G)$ and $\Gamma^-\left(G^C\right)$ are both in $\mathscr{F}$. Then clearly $\mathscr{G}$ is closed with respect to complements. If $G \in \mathscr{G}$ is G^C? Is $\Gamma^-\left(G^C\right)$ and $\Gamma^-\left(\left(G^C\right)^C\right)$ in $\mathscr{F}$? This is just the definition of what it means to be in $\mathscr{G}$. Also if you have $\{G_i\} \subseteq \mathscr{G}$, Then

$$\Gamma^-(\cup_i G_i) = \cup_i \Gamma^-(G_i) \in \mathscr{F}$$

and so $\mathscr{G}$ is closed with respect to countable unions. Hence $\mathscr{G}$ must contain the Borel sets because it is a σ algebra and the closed sets are in $\mathscr{G}$. Thus $\Gamma^-(G) \in \mathscr{F}$ whenever G is Borel.

9.15.3 Kuratowski's Theorem

Also there is a useful corollary from Theorem 9.15.2 and Proposition 9.15.5.

Corollary 9.15.6 *Let $K(\omega)$ be a compact subset of a separable metric space X and suppose $\{u_j(\omega)\}_{j=1}^{\infty} \subseteq K(\omega)$ with each $\omega \to u_j(\omega)$ measurable into X. Then there exists $u(\omega) \in K(\omega)$ such that $\omega \to u(\omega)$ is measurable into X and a subsequence $n(\omega)$ depending on ω such that $\lim_{n(\omega)\to\infty} u_{n(\omega)}(\omega) = u(\omega)$.*

Proof: Define $\Gamma_n(\omega) = \overline{\cup_{k\geq n} u_k(\omega)}$.This is a nonempty compact subset of $K(\omega) \subseteq X$. I claim that $\omega \to \Gamma_n(\omega)$ is a measurable multifunction into X. It is necessary to show that $\Gamma_n^-(O)$ defined as $\{\omega : \Gamma_n(\omega) \cap O \neq \emptyset\}$ is measurable whenever O is open in X. For $\omega \in \Gamma_n^-(O)$ it means that some $u_k(\omega) \in O, k \geq n$. Thus $\Gamma_n^-(O) = \cup_{k\geq n} u_k^{-1}(O)$ and this is measurable by the assumption that each u_k is. Since $\Gamma_n^-(\omega)$ is compact, it is also strongly measurable by Proposition 9.15.5, meaning that $\Gamma^-(H)$ is measurable whenever H is closed. Now, let $\Gamma(\omega)$ be defined as $\Gamma(\omega) \equiv \cap_n \Gamma_n(\omega)$ and then for H closed, $\Gamma^-(H)$ is nonempty if and only if $\Gamma_n^-(H)$ is nonempty for each n and $\Gamma^-(H) = \cap_n \Gamma_n^-(H)$ and each set in the intersection is measurable, so this shows that $\omega \to \Gamma(\omega)$ is also (strongly) measurable. Therefore, it has a measurable selection $u(\omega)$. It follows from the definition of $\Gamma(\omega)$ that there exists a subsequence $n(\omega)$ such that $u(\omega) = \lim_{n(\omega)\to\infty} u_{n(\omega)}(\omega)$ ■

This corollary makes possible a fairly short proof of the very amazing and enormously significant Kuratowski theorem [35] which gives measurability of maximums of Carathéodory functions.

Definition 9.15.7 *The functions $f : \Omega \times E \to \mathbb{R}$ in which $f(\cdot, \omega)$ is continuous and $\omega \to f(x, \omega)$ is measurable are called Caratheodory functions.*

Now here is the Kuratowski theorem.

Theorem 9.15.8 *Let E be a compact metric space and let $(\Omega, \mathscr{F})$ be a measure space. Suppose $\psi : E \times \Omega \to \mathbb{R}$ has the property that $x \to \psi(x, \omega)$ is continuous and $\omega \to \psi(x, \omega)$ is measurable. Then there exists a measurable function f having values in E such that $\psi(f(\omega), \omega) = \max_{x\in E} \psi(x, \omega)$. Furthermore, $\omega \to \psi(f(\omega), \omega)$ is measurable.*

Proof: Let $C = \{e_i\}_{i=1}^{\infty}$ be a countable dense subset of E. For example, take the union of $1/2^n$ nets for all n. Let $C_n \equiv \{e_1, ..., e_n\}$. Let $\omega \to f_n(\omega)$ be measurable and satisfy $\psi(f_n(\omega), \omega) = \sup_{x \in C_n} \psi(x, \omega)$. This is easily done as follows. Let

$$B_k \equiv \left\{\omega : \psi(e_k, \omega) \geq \psi(e_j, \omega) \text{ for all } j \neq k\right\}.$$

Then let $A_1 \equiv B_1$ and if $A_1, ..., A_k$ have been chosen, let $A_{k+1} \equiv B_{k+1} \setminus \left(\cup_{j=1}^{k} B_k\right)$. Thus each A_k is measurable and you let $f_n(\omega) \equiv e_k$ for $\omega \in A_k$. Using Corollary 9.15.6, there is measurable $f(\omega)$ and a subsequence $n(\omega) \geq n$ such that $f_{n(\omega)}(\omega) \to f(\omega)$. Then by continuity, $\psi(f(\omega), \omega) = \lim_{n(\omega) \to \infty} \psi\left(f_{n(\omega)}(\omega), \omega\right)$ and this is an increasing sequence in this limit. Hence $\psi(f(\omega), \omega) \geq \sup_{x \in C_n} \psi(x, \omega)$ for each n and so $\psi(f(\omega), \omega) \geq \sup_{x \in C} \psi(x, \omega) = \sup_{x \in E} \psi(x, \omega)$. Since f is measurable, it is the limit of a sequence $\{g_n(\omega)\}$ such that g_n has finitely many values occuring on measurable sets, Theorem 9.1.7. Hence, by continuity, $\psi(f(\omega), \omega) = \lim_{n \to \infty} \psi(g_n(\omega), \omega)$ and since $\omega \to \psi(g_n(\omega), \omega)$ is measurable, so is $\psi(f(\omega), \omega)$. ■

One can generalize fairly easily. It is the same argument but carrying around more ω.

Theorem 9.15.9 *Let $E(\omega)$ be a compact metric space in a separable metric space (X, d) and suppose that $\omega \to E(\omega)$ is a measurable multifunction where $(\Omega, \mathscr{F})$ be a measure space. Suppose $\psi_\omega : E(\omega) \times \Omega \to \mathbb{R}$ has the property that $x \to \psi_\omega(x, \omega)$ is continuous and $\omega \to \psi_\omega(x(\omega), \omega)$ is measurable if $x(\omega) \in E(\omega)$ and $\omega \to x(\omega)$ is measurable ($x(\omega)$ a measurable selection of $E(\omega)$). Then there exists a measurable function f with $f(\omega) \in E(\omega)$ such that $\psi_\omega(f(\omega), \omega) = \max_{x \in E(\omega)} \psi_\omega(x, \omega)$. Furthermore, $\omega \to \psi_\omega(f(\omega), \omega)$ is measurable.*

Proof: Let $C(\omega) = \{e_i(\omega)\}_{i=1}^{\infty}$ be a countable dense subset of $E(\omega)$ with each $e_i(\omega)$ measurable. This countable dense subset exists by Theorem 9.15.2. Let

$$C_n(\omega) \equiv \{e_1(\omega), ..., e_n(\omega)\}.$$

Let $\omega \to f_n(\omega)$ be measurable and satisfy

$$\psi_\omega(f_n(\omega), \omega) = \sup_{x \in C_n(\omega)} \psi(x, \omega).$$

This is easily done as follows. Let

$$B_k \equiv \left\{\omega : \psi_\omega(e_k(\omega), \omega) \geq \psi_\omega(e_j(\omega), \omega) \text{ for all } j \neq k\right\}.$$

Then let $A_1 \equiv B_1$ and if $A_1, ..., A_k$ have been chosen, let $A_{k+1} \equiv B_{k+1} \setminus \left(\cup_{j=1}^{k} B_k\right)$. Thus each A_k is measurable, and you let $f_n(\omega) \equiv e_k(\omega)$ for $\omega \in A_k$, so $f_n(\omega) \in E(\omega)$ and f_n is measurable. Using Corollary 9.15.6, there is measurable $f(\omega)$ and a subsequence $n(\omega) \geq n$ such that $f_{n(\omega)}(\omega) \to f(\omega)$. Then by continuity,

$$\psi_\omega(f(\omega), \omega) = \lim_{n(\omega) \to \infty} \psi_\omega\left(f_{n(\omega)}(\omega), \omega\right)$$

and this is an increasing sequence in this limit. Hence

$$\psi_\omega(f(\omega), \omega) \geq \sup_{x \in C_n(\omega)} \psi_\omega(x, \omega)$$

for each n and so

$$\psi_\omega(f(\omega),\omega) \geq \sup_{x\in C(\omega)} \psi_\omega(x,\omega) = \sup_{x\in E(\omega)} \psi_\omega(x,\omega).$$

Since f is measurable, it follows by assumption, that $\omega \to \psi_\omega(f(\omega),\omega)$ is measurable. ■

Note the following: If you have the simpler situation where $\psi(x,\omega)$ defined on $X\times\Omega$ with $x\to\psi(x,\omega)$ continuous and $\omega\to\psi(x,\omega)$ measurable but $E(\omega)$ a compact measurable multifunction as above, then the conditions will hold because you would have $\omega\to\psi(x(\omega),\omega)$ is measurable if $x(\omega)$ is. Indeed, $x(\omega)$ is the limit of a sequence $\{x_n(\omega)\}$ such that x_n has finitely many values on measurable sets, Theorem 9.1.7. Hence, by continuity, $\psi(x(\omega),\omega) = \lim_{n\to\infty}\psi(x_n(\omega),\omega)$ and since $\omega\to\psi(x_n(\omega),\omega)$ is measurable, so is $\psi(x(\omega),\omega)$.

9.15.4 Measurability of Fixed Points

As an interesting application is a consideration of the existence of measurable Brouwer fixed points. This is really quite amazing since Brouwer fixed points are not obtained as the limit of a sequence of iterates although the above Sperner's lemma algorithm provides an algorithm for finding one. This is a very nice application of the marvelous Kuratowski theorem. It is possible to get this result directly from Corollary 9.15.6 applied to the Sperner's lemma method of proving the Brouwer fixed point theorem. We sent in a paper once which did this and we thought the result was amazing. However, I had forgotten about Kuratowski's theorem which makes this very easy. Fortunately, the referee knew this theorem.

Theorem 9.15.10 *Let K be a closed convex bounded subset of $\mathbb{R}^p$. Let*

$$\boldsymbol{x}\to\boldsymbol{f}(\boldsymbol{x},\omega): K\to K$$

be continuous for each ω and $\omega\to\boldsymbol{f}(\boldsymbol{x},\omega)$ is measurable, meaning inverse images of sets open in K are in $\mathscr{F}$ where $(\Omega,\mathscr{F})$ is a measurable space. Then there exists $\boldsymbol{x}(\omega)\in K$ such that $\omega\to\boldsymbol{x}(\omega)$ is measurable and $\boldsymbol{f}(\boldsymbol{x}(\omega),\omega)=\boldsymbol{x}(\omega)$.

Proof: Simply consider $E=K$ and $\psi(\boldsymbol{x},\omega)\equiv -|\boldsymbol{x}-\boldsymbol{f}(\boldsymbol{x},\omega)|$. It has a maximum $\boldsymbol{x}(\omega)$ for each ω thanks to continuity of $\boldsymbol{f}(\cdot,\omega)$. Thanks to the Brouwer fixed point theorem, this $\boldsymbol{x}(\omega)$ must be a fixed point. By the above Kuratowski theorem, one of these $\boldsymbol{x}(\omega)$ is measurable. Obviously, by continuity of $\boldsymbol{f}(\cdot,\omega)$, $\omega\to\boldsymbol{f}(\boldsymbol{x}(\omega),\omega)$ is measurable. ■

If desired, you can extend this to the case where $K(\omega)$ is a measurable multifunction.

9.15.5 Other Measurability Considerations

Here are some other general considerations about measurable multifunctions. The first has to do with getting a new measurable multifunction from old ones and the second has to do with measurability of ε nets. These are technical results which are sometimes useful, for example, if you want to generalize to the Schauder fixed point theorem.

Lemma 9.15.11 *Suppose $f: K(\omega)\times\Omega\to X, K\subseteq X$. Here X is Polish space, separable complete metric space, and $(\Omega,\mathscr{F})$ is a measurable space. Also $\omega\to K(\omega)$ is a measurable multifunction as in Theorem 9.15.2. Also suppose*

1. *$\omega\to f(x,\omega)$ is measurable and $x\to f(x,\omega)$ is continuous.*

2. $\mathscr{K}(\omega) \equiv f(K(\omega), \omega)$.

Then you can conclude that $\omega \to \mathscr{K}(\omega)$ is a measurable multifunction. If $\mathscr{K}(\omega)$ is compact, then it is also strongly measurable.

Proof: Let $\{x_n(\omega)\}$ be a countable dense subset of $K(\omega)$, each x_n measurable. Then if U is open,

$$\{\omega : \mathscr{K}(\omega) \cap U \neq \emptyset\} = \cup_{n=1}^{\infty} f(x_n(\cdot), \cdot)^{-1}(U) \tag{9.26}$$

and each of the sets in the union is measurable. The latter claim follows from the continuity of $f(\cdot, \omega)$. If $x(\omega)$ is measurable, then we can express it as the limit of functions s_n which have finitely many values on measurable sets for which $\omega \to f(s_n(\omega), \omega)$ is clearly measurable. Then $f(x(\omega), \omega)$ is the limit of $f(s_n(\omega), \omega)$. The reason for the equality in 9.26 is as follows. It is clear that the right side is contained in the left. Now if $\mathscr{K}(\omega) \cap U \neq \emptyset$, then by definition, $f(x, \omega) \in U$ for some $x \in K(\omega)$ but then by continuity, $f(x_n(\omega), \omega) \in U$ also for some $x_n(\omega)$ close to x. Thus the two sets are actually equal. Thus $\omega \to \mathscr{K}(\omega)$ is measurable. If $\mathscr{K}(\omega)$ has compact values it will be strongly measurable as discussed in Proposition 9.15.5. ■

There is also the following general result about the existence of a measurable ε net. This is formulated in Banach space because it is convenient to add. A Banch space is just a complete normed vector space. It could also be formulated in Polish space with a little more difficulty. One just defines things a little differently.

Proposition 9.15.12 *Let $\omega \to \mathscr{K}(\omega)$ be a measurable multifunction where $\mathscr{K}(\omega)$ is a pre compact set. Recall this means its closure is compact. Thus $\mathscr{K}(\omega)$ must have an ε net for each $\varepsilon > 0$. Then for each $\varepsilon > 0$, there exists $N(\omega)$ and measurable functions $y_j, j = 1, 2, \cdots, N(\omega)$, $y_j(\omega) \in \mathscr{K}(\omega)$, such that $\cup_{j=1}^{N(\omega)} B(y_j(\omega), \varepsilon) \supseteq \mathscr{K}(\omega)$ for each ω. Also $\omega \to N(\omega)$ is measurable.*

Proof: Suppose that $\omega \to \mathscr{K}(\omega)$ is a measurable multifunction having compact values in X a Banach space. Let $\{\sigma_n(\omega)\}$ be the measurable selections such that for each $\omega, \{\sigma_n(\omega)\}_{n=1}^{\infty}$ is dense in $\mathscr{K}(\omega)$. Let $y_1(\omega) \equiv \sigma_1(\omega)$. Now let $2(\omega)$ be the first index larger than 1 such that $\|\sigma_{2(\omega)}(\omega) - \sigma_1(\omega)\| > \frac{\varepsilon}{2}$. Thus $2(\omega) = k$ on the measurable set

$$\left\{\omega \in \Omega : \|\sigma_k(\omega) - \sigma_1(\omega)\| > \frac{\varepsilon}{2}\right\} \cap \left\{\omega \in \Omega : \cap_{j=1}^{k-1} \|\sigma_j(\omega) - \sigma_1(\omega)\| \leq \frac{\varepsilon}{2}\right\}$$

Suppose $1(\omega), 2(\omega), \cdots, (m-1)(\omega)$ have been chosen such that this is a strictly increasing sequence for each ω, each is a measurable function, and for $i, j \leq m-1$,

$$\|\sigma_{i(\omega)}(\omega) - \sigma_{j(\omega)}(\omega)\| > \frac{\varepsilon}{2}.$$

Each $\omega \to \sigma_{i(\omega)}(\omega)$ is measurable since it equals $\sum_{k=1}^{\infty} \mathscr{X}_{[i(\omega)=k]}(\omega)\sigma_k(\omega)$. Then $m(\omega)$ will be the first index larger than $(m-1)(\omega)$ such that

$$\|\sigma_{m(\omega)}(\omega) - \sigma_{j(\omega)}(\omega)\| > \frac{\varepsilon}{2}$$

for all $j(\omega) < m(\omega)$. Thus $\omega \to m(\omega)$ is also measurable because it equals k on the measurable set

$$\left(\cap \left\{\omega : \|\sigma_k(\omega) - \sigma_{j(\omega)}(\omega)\| > \frac{\varepsilon}{2},\ j \leq m-1\right\}\right) \cap \{\omega : (m-1)(\omega) < k\}$$

$$\cap\left(\cup\left\{\omega : \left\|\sigma_{k-1}(\omega)-\sigma_{j(\omega)}(\omega)\right\| \leq \frac{\varepsilon}{2},\ j \leq m-1\right\}\right)$$

The top line says that it does what is wanted and the second says it is the first after $(m-1)(\omega)$ which does so.

Since $\mathscr{K}(\omega)$ is a pre compact set, it follows that the above measurable set will be empty for all $m(\omega)$ sufficiently large called $N(\omega)$, also a measurable function, and so the process ends. Let $y_i(\omega) \equiv \sigma_{i(\omega)}(\omega)$. Then this gives the desired measurable ε net. The fact that

$$\cup_{i=1}^{N(\omega)} B(y_i(\omega),\varepsilon) \supseteq \mathscr{K}(\omega)$$

follows because if there exists $z \in \mathscr{K}(\omega) \setminus \left(\cup_{i=1}^{N(\omega)} B(y_i(\omega),\varepsilon)\right)$, then $B\left(z,\frac{\varepsilon}{2}\right)$ would have empty intersection with all of the balls $B\left(y_i(\omega),\frac{\varepsilon}{3}\right)$ and by density of the $\sigma_i(\omega)$ in $\mathscr{K}(\omega)$, there would be some $\sigma_l(\omega)$ contained in $B\left(z,\frac{\varepsilon}{3}\right)$ for arbitrarily large l and so the process would not have ended as shown above. ■

9.16 Exercises

1. Using some form of Kuratowski's theorem show the following: Let $K(\omega)$ be a closed convex bounded subset of $\mathbb{R}^n$ where $\omega \to K(\omega)$ is a measurable multifunction. Let $\boldsymbol{x} \to \boldsymbol{f}(\boldsymbol{x},\omega) : K(\omega) \to K(\omega)$ be continuous for each ω and $\omega \to \boldsymbol{f}(\boldsymbol{x},\omega)$ is measurable, meaning inverse images of sets open in $\mathbb{R}^n$ are in $\mathscr{F}$ where $(\Omega,\mathscr{F})$ is a measurable space. Then there exists $\boldsymbol{x}(\omega) \in K(\omega)$ such that $\omega \to \boldsymbol{x}(\omega)$ is measurable and $\boldsymbol{f}(\boldsymbol{x}(\omega),\omega) = \boldsymbol{x}(\omega)$.

2. If you have $K(\omega)$ a closed convex nonempty set in $\mathbb{R}^n$ and also $\omega \to K(\omega)$ is a measurable multifunction, show $\omega \to P_{K(\omega)}\boldsymbol{x}$ is measurable where $P_{K(\omega)}$ is the projection map which gives the closest point in $K(\omega)$. Consider Corollary 6.3.2 on Page 165 or Theorem 11.11.9 and Problem 10 on Page 152 to see the use of this projection map. Also you may want to use Theorem 9.15.2 involving the countable dense subset of $K(\omega)$ consisting of measurable functions.

3. Let $\omega \to K(\omega)$ be a measurable multifunction in $\mathbb{R}^p$ and let $K(\omega)$ be convex, closed, and compact for each ω. Let $A(\cdot,\omega) : K(\omega) \to \mathbb{R}^p$ be continuous and $\omega \to A(\boldsymbol{x},\omega)$ be measurable. Then if $\omega \to \boldsymbol{y}(\omega)$ is measurable, there exists measurable $\omega \to \boldsymbol{x}(\omega)$ such that for all $\boldsymbol{z} \in K(\omega)$,

$$(\boldsymbol{y}(\omega) - A(\boldsymbol{x}(\omega),\omega), \boldsymbol{z}(\omega) - \boldsymbol{x}(\omega)) \leq 0$$

   This is a measurable version of Browder's lemma, a very important result in nonlinear analysis. **Hint:** You want to have for each ω,

$$P_{K(\omega)}(\boldsymbol{y}(\omega) - A(\boldsymbol{x},\omega) + \boldsymbol{x}) = \boldsymbol{x}$$

   Use Problem 2 and the measurability of Brouwer fixed points discussed above.

4. In the situation of the above problem, suppose also that $\lim_{|\boldsymbol{x}|\to\infty} \frac{(A(\boldsymbol{x},\omega),\boldsymbol{x})}{|\boldsymbol{x}|} = \infty$ Show that there exists measurable $\boldsymbol{x}(\omega)$ such that $A(\boldsymbol{x}(\omega),\omega) = \boldsymbol{y}(\omega)$. **Hint:** Let $\boldsymbol{x}_n(\omega)$ be the solution of Problem 3 in which $K_n = \overline{B(\boldsymbol{0},n)}$. Show that these are bounded for each ω. Then use Corollary 9.15.6 to get $\boldsymbol{x}(\omega)$, a suitable limit such that $A(\boldsymbol{x}(\omega),\omega) = \boldsymbol{y}(\omega)$.

Chapter 10

The Abstract Lebesgue Integral

The general Lebesgue integral requires a measure space, $(\Omega, \mathscr{F}, \mu)$ and, to begin with, a nonnegative measurable function. I will use Lemma 2.5.3 about interchanging two supremums frequently. Also, I will use the observation that if $\{a_n\}$ is an increasing sequence of points of $[0,\infty]$, then $\sup_n a_n = \lim_{n\to\infty} a_n$ which is obvious from the definition of sup.

10.1 Nonnegative Measurable Functions

10.1.1 Riemann Integrals For Decreasing Functions

First of all, the notation $[g < f]$ means $\{\omega \in \Omega : g(\omega) < f(\omega)\}$ with other variants of this notation being similar. Also, the convention, $0 \cdot \infty = 0$ will be used to simplify the presentation whenever it is convenient to do so. The notation $a \wedge b$ means the minimum of a and b.

Definition 10.1.1 *Let $f : [a,b] \to [0,\infty]$ be decreasing. Note that ∞ is a possible value. Define*

$$\int_a^b f(\lambda)\, d\lambda \equiv \lim_{M\to\infty} \int_a^b M \wedge f(\lambda)\, d\lambda = \sup_M \int_a^b M \wedge f(\lambda)\, d\lambda$$

where $a \wedge b$ means the minimum of a and b. Note that for f bounded,

$$\sup_M \int_a^b M \wedge f(\lambda)\, d\lambda = \int_a^b f(\lambda)\, d\lambda$$

where the integral on the right is the usual Riemann integral because eventually $M > f$. For f a nonnegative decreasing function defined on $[0,\infty)$,

$$\int_0^\infty f d\lambda \equiv \lim_{R\to\infty} \int_0^R f d\lambda = \sup_{R>1} \int_0^R f d\lambda = \sup_R \sup_{M>0} \int_0^R f \wedge M d\lambda$$

Since decreasing bounded functions are Riemann integrable, the above definition is well defined. Now here is an obvious property.

Lemma 10.1.2 *Let f be a decreasing nonnegative function defined on an interval $[a,b]$. Then if $[a,b] = \cup_{k=1}^m I_k$ where $I_k \equiv [a_k, b_k]$ and the intervals I_k are non overlapping, it follows*

$$\int_a^b f d\lambda = \sum_{k=1}^m \int_{a_k}^{b_k} f d\lambda.$$

Proof: This follows from the computation,

$$\begin{aligned} \int_a^b f d\lambda &\equiv \lim_{M\to\infty} \int_a^b f \wedge M d\lambda \\ &= \lim_{M\to\infty} \sum_{k=1}^m \int_{a_k}^{b_k} f \wedge M d\lambda = \sum_{k=1}^m \int_{a_k}^{b_k} f d\lambda \end{aligned}$$

Note both sides could equal $+\infty$. ■

In all considerations below, we assume h is fairly small, certainly much smaller than R. Thus $R - h > 0$.

Lemma 10.1.3 *Let g be a decreasing nonnegative function defined on an interval $[0,R]$. Then*

$$\int_0^R g\wedge M d\lambda = \sup_{h>0}\sum_{i=1}^{m(R,h)} g(ih)\,h$$

where $m(h,R)\in\mathbb{N}$ satisfies $R-h<hm(h,R)\leq R$.

Proof: Since $g\wedge M$ is a decreasing bounded function the lower sums converge to the integral as $h\to 0$. Thus

$$\int_0^R g\wedge M d\lambda = \lim_{h\to 0}\left(\sum_{i=1}^{m(R,h)}(g(ih)\wedge M)\,h + (g(R)\wedge M)(R-hm(h,R))\right)$$

Now the last term in the above is no more than Mh and so the above is

$$\lim_{h\to 0}\left(\sum_{i=1}^{m(R,h)}(g(ih)\wedge M)\,h\right) = \sup_{h>0}\left(\sum_{i=1}^{m(R,h)}(g(ih)\wedge M)\,h\right). \blacksquare$$

10.1.2 The Lebesgue Integral for Nonnegative Functions

Here is the definition of the Lebesgue integral of a function which is measurable and has values in $[0,\infty]$.

Definition 10.1.4 *Let $(\Omega,\mathcal{F},\mu)$ be a measure space and suppose $f:\Omega\to[0,\infty]$ is measurable. Then define*

$$\int f d\mu \equiv \int_0^\infty \mu([f>\lambda])\,d\lambda$$

which makes sense because $\lambda\to\mu([f>\lambda])$ is nonnegative and decreasing.

Note that if $f\leq g$, then $\int f d\mu\leq\int g d\mu$ because $\mu([f>\lambda])\leq\mu([g>\lambda])$.
For convenience $\sum_{i=1}^0 a_i\equiv 0$.

Lemma 10.1.5 *In the situation of the above definition,*

$$\int f d\mu = \sup_{h>0}\sum_{i=1}^\infty \mu([f>hi])\,h$$

Proof: Let $m(h,R)\in\mathbb{N}$ satisfy $R-h<hm(h,R)\leq R$. Then

$$\lim_{R\to\infty} m(h,R)=\infty$$

and so from Lemma 10.1.3,

$$\begin{aligned}\int f d\mu &\equiv \int_0^\infty \mu([f>\lambda])\,d\lambda = \sup_M\sup_R\int_0^R \mu([f>\lambda])\wedge M d\lambda\\ &= \sup_M\sup_{R>0}\sup_{h>0}\sum_{k=1}^{m(h,R)}(\mu([f>kh])\wedge M)\,h\end{aligned}$$

Hence, switching the order of the sups, this equals

$$\sup_{R>0}\sup_{h>0}\sup_{M}\sum_{k=1}^{m(h,R)}\left(\mu\left([f>kh]\right)\wedge M\right)h=\sup_{R>0}\sup_{h>0}\lim_{M\to\infty}\sum_{k=1}^{m(h,R)}\left(\mu\left([f>kh]\right)\wedge M\right)h$$

$$=\sup_{h>0}\sup_{R}\sum_{k=1}^{m(R,h)}\left(\mu\left([f>kh]\right)\right)h=\sup_{h>0}\sum_{k=1}^{\infty}\left(\mu\left([f>kh]\right)\right)h.\ \blacksquare$$

10.2 Nonnegative Simple Functions

To begin with, here is a useful lemma.

Lemma 10.2.1 *If $f(\lambda)=0$ for all $\lambda>a$, where f is a decreasing nonnegative function, then*

$$\int_0^\infty f(\lambda)\,d\lambda=\int_0^a f(\lambda)\,d\lambda.$$

Proof: From the definition,

$$\begin{aligned}\int_0^\infty f(\lambda)\,d\lambda &= \lim_{R\to\infty}\int_0^R f(\lambda)\,d\lambda=\sup_{R>1}\int_0^R f(\lambda)\,d\lambda\\ &= \sup_{R>1}\sup_{M}\int_0^R f(\lambda)\wedge M\,d\lambda\\ &= \sup_{M}\sup_{R>1}\int_0^R f(\lambda)\wedge M\,d\lambda\\ &= \sup_{M}\sup_{R>1}\int_0^a f(\lambda)\wedge M\,d\lambda\\ &= \sup_{M}\int_0^a f(\lambda)\wedge M\,d\lambda\equiv\int_0^a f(\lambda)\,d\lambda.\ \blacksquare\end{aligned}$$

Now the Lebesgue integral for a nonnegative function has been defined, what does it do to a nonnegative simple function? Recall a nonnegative simple function is one which has finitely many nonnegative real values which it assumes on measurable sets. Thus a simple function can be written in the form

$$s(\omega)=\sum_{i=1}^n c_i\mathcal{X}_{E_i}(\omega)$$

where the c_i are each nonnegative, the distinct values of s.

Lemma 10.2.2 *Let $s(\omega)=\sum_{i=1}^p a_i\mathcal{X}_{E_i}(\omega)$ be a nonnegative simple function where the E_i are distinct but the a_i might not be. Thus the values of s are the a_i. Then*

$$\int s\,d\mu=\sum_{i=1}^p a_i\mu(E_i).\tag{10.1}$$

Proof: Without loss of generality, assume $0\equiv a_0<a_1\le a_2\le\cdots\le a_p$ and that $\mu(E_i)<\infty, i>0$. Here is why. If $\mu(E_i)=\infty$, then letting $a\in(a_{i-1},a_i)$, by Lemma 10.2.1, the left

side is

$$\begin{aligned}\int_0^{a_p} \mu([s>\lambda])\,d\lambda &\geq \int_{a_0}^{a_i} \mu([s>\lambda])\,d\lambda \\ &\equiv \sup_M \int_0^{a_i} \mu([s>\lambda]) \wedge M d\lambda \\ &\geq \sup_M M a_i = \infty\end{aligned}$$

and so both sides of 10.1 are equal to ∞. Thus it can be assumed for each $i, \mu(E_i) < \infty$. Then it follows from Lemma 10.2.1 and Lemma 10.1.2,

$$\int_0^{\infty} \mu([s>\lambda])\,d\lambda = \int_0^{a_p} \mu([s>\lambda])\,d\lambda = \sum_{k=1}^{p} \int_{a_{k-1}}^{a_k} \mu([s>\lambda])\,d\lambda$$

$$= \sum_{k=1}^{p} (a_k - a_{k-1}) \sum_{i=k}^{p} \mu(E_i) = \sum_{i=1}^{p} \mu(E_i) \sum_{k=1}^{i} (a_k - a_{k-1}) = \sum_{i=1}^{p} a_i \mu(E_i) \blacksquare$$

Note that this is the same result as in Problem 12 on Page 274 but here there is no question about the definition of the integral of a simple function being well defined.

Lemma 10.2.3 *If $a, b \geq 0$ and if s and t are nonnegative simple functions, then*

$$\int as + bt\,d\mu = a \int s\,d\mu + b \int t\,d\mu.$$

Proof: Let

$$s(\omega) = \sum_{i=1}^{n} \alpha_i \mathcal{X}_{A_i}(\omega),\ t(\omega) = \sum_{i=1}^{m} \beta_j \mathcal{X}_{B_j}(\omega)$$

where α_i are the distinct values of s and the β_j are the distinct values of t. Clearly $as + bt$ is a nonnegative simple function because it has finitely many values on measurable sets. In fact,

$$(as+bt)(\omega) = \sum_{j=1}^{m} \sum_{i=1}^{n} (a\alpha_i + b\beta_j) \mathcal{X}_{A_i \cap B_j}(\omega)$$

where the sets $A_i \cap B_j$ are disjoint and measurable. By Lemma 10.2.2,

$$\begin{aligned}\int as+bt\,d\mu &= \sum_{j=1}^{m} \sum_{i=1}^{n} (a\alpha_i + b\beta_j) \mu(A_i \cap B_j) \\ &= \sum_{i=1}^{n} a \sum_{j=1}^{m} \alpha_i \mu(A_i \cap B_j) + b \sum_{j=1}^{m} \sum_{i=1}^{n} \beta_j \mu(A_i \cap B_j) \\ &= a \sum_{i=1}^{n} \alpha_i \mu(A_i) + b \sum_{j=1}^{m} \beta_j \mu(B_j) \\ &= a \int s\,d\mu + b \int t\,d\mu. \blacksquare\end{aligned}$$

10.3 The Monotone Convergence Theorem

The following is called the monotone convergence theorem. This theorem and related convergence theorems are the reason for using the Lebesgue integral. If $\lim_{n\to\infty} f_n(\omega) = f(\omega)$ and f_n is increasing in n, then clearly f is also measurable because

$$f^{-1}((a,\infty]) = \cup_{k=1}^{\infty} f_k^{-1}((a,\infty]) \in \mathscr{F}$$

For a different approach to this, see Problem 12 on Page 274.

Theorem 10.3.1 *(Monotone Convergence theorem) Suppose that the function f has all values in $[0,\infty]$ and suppose $\{f_n\}$ is a sequence of nonnegative measurable functions having values in $[0,\infty]$ and satisfying*

$$\lim_{n\to\infty} f_n(\omega) = f(\omega) \text{ for each } \omega.$$

$$\cdots f_n(\omega) \le f_{n+1}(\omega) \cdots$$

Then f is measurable and

$$\int f d\mu = \lim_{n\to\infty} \int f_n d\mu.$$

Proof: By Lemma 10.1.5

$$\lim_{n\to\infty} \int f_n d\mu = \sup_n \int f_n d\mu$$

$$\begin{aligned} &= \sup_n \sup_{h>0} \sum_{k=1}^{\infty} \mu([f_n > kh])\, h = \sup_{h>0} \sup_N \sup_n \sum_{k=1}^{N} \mu([f_n > kh])\, h \\ &= \sup_{h>0} \sup_N \sum_{k=1}^{N} \mu([f > kh])\, h = \sup_{h>0} \sum_{k=1}^{\infty} \mu([f > kh])\, h = \int f d\mu. \ \blacksquare \end{aligned}$$

Note how it was important to have $\int_0^\infty [f > \lambda]\, d\lambda$ in the definition of the integral and **not** $[f \ge \lambda]$. You need to have $[f_n > kh] \uparrow [f > kh]$ so $\mu([f_n > kh]) \to \mu([f > kh])$. To illustrate what goes wrong without the Lebesgue integral, consider the following example.

Example 10.3.2 *Let $\{r_n\}$ denote the rational numbers in $[0,1]$ and let*

$$f_n(t) \equiv \begin{cases} 1 \text{ if } t \notin \{r_1, \cdots, r_n\} \\ 0 \text{ otherwise} \end{cases}$$

Then $f_n(t) \uparrow f(t)$ where f is the function which is one on the rationals and zero on the irrationals. Each f_n is Riemann integrable (why?) but f is not Riemann integrable because it is everywhere discontinuous. Also, there is a gap between all upper sums and lower sums. Therefore, you can't write $\int f dx = \lim_{n\to\infty} \int f_n dx$.

An observation which is typically true related to this type of example is this. If you can choose your functions, you don't need the Lebesgue integral. The Riemann Darboux integral is just fine. It is when you can't choose your functions and they come to you as pointwise limits that you really need the superior Lebesgue integral or at least something more general than the Riemann integral. The Riemann integral is entirely adequate for evaluating the seemingly endless lists of boring problems found in calculus books. It is shown later that the two integrals coincide when the Lebesgue integral is taken with respect to Lebesgue measure and the function being integrated is continuous.

10.4 Other Definitions

To review and summarize the above, if $f \geq 0$ is measurable,

$$\int f d\mu \equiv \int_0^\infty \mu([f > \lambda]) d\lambda \tag{10.2}$$

another way to get the same thing for $\int f d\mu$ is to take an increasing sequence of nonnegative simple functions, $\{s_n\}$ with $s_n(\omega) \to f(\omega)$ and then by monotone convergence theorem, $\int f d\mu = \lim_{n\to\infty} \int s_n$ where if $s_n(\omega) = \sum_{j=1}^m c_i \mathcal{X}_{E_i}(\omega)$, $\int s_n d\mu = \sum_{i=1}^m c_i \mu(E_i)$. Similarly this also shows that for such nonnegative measurable function,

$$\int f d\mu = \sup\left\{\int s : 0 \leq s \leq f,\ s \text{ simple}\right\}.$$

Here is an equivalent definition of the integral of a nonnegative measurable function. The fact it is well defined has been discussed above.

Definition 10.4.1 *For s a nonnegative simple function,*

$$s(\omega) = \sum_{k=1}^n c_k \mathcal{X}_{E_k}(\omega), \int s = \sum_{k=1}^n c_k \mu(E_k).$$

For f a nonnegative measurable function,

$$\int f d\mu = \sup\left\{\int s : 0 \leq s \leq f,\ s \textit{ simple}\right\}.$$

10.5 Fatou's Lemma

The next theorem, known as Fatou's lemma is another important theorem which justifies the use of the Lebesgue integral.

Theorem 10.5.1 *(Fatou's lemma) Let f_n be a nonnegative measurable function. Let $g(\omega) = \liminf_{n\to\infty} f_n(\omega)$. Then g is measurable and*

$$\int g d\mu \leq \liminf_{n\to\infty} \int f_n d\mu.$$

In other words,

$$\int \left(\liminf_{n\to\infty} f_n\right) d\mu \leq \liminf_{n\to\infty} \int f_n d\mu$$

Proof: Let $g_n(\omega) = \inf\{f_k(\omega) : k \geq n\}$. Then

$$g_n^{-1}([a,\infty]) = \cap_{k=n}^\infty f_k^{-1}([a,\infty]) = \left(\cup_{k=n}^\infty f_k^{-1}([a,\infty])^C\right)^C \in \mathcal{F}.$$

Thus g_n is measurable by Lemma 9.1.4. Also $g(\omega) = \lim_{n\to\infty} g_n(\omega)$ so g is measurable because it is the pointwise limit of measurable functions. Now the functions g_n form an increasing sequence of nonnegative measurable functions so the monotone convergence theorem applies. This yields

$$\int g d\mu = \lim_{n\to\infty} \int g_n d\mu \leq \liminf_{n\to\infty} \int f_n d\mu.$$

The last inequality holding because $\int g_n d\mu \leq \int f_n d\mu$. (Note that it is not known whether $\lim_{n\to\infty} \int f_n d\mu$ exists.) ■

10.6 The Integral's Righteous Algebraic Desires

The monotone convergence theorem shows the integral wants to be linear. This is the essential content of the next theorem.

Theorem 10.6.1 *Let f,g be nonnegative measurable functions and let a,b be nonnegative numbers. Then $af+bg$ is measurable and*

$$\int (af+bg)\,d\mu = a\int f\,d\mu + b\int g\,d\mu. \tag{10.3}$$

Proof: By Theorem 9.1.6 on Page 244 there exist increasing sequences of nonnegative simple functions, $s_n \to f$ and $t_n \to g$. Then $af+bg$, being the pointwise limit of the simple functions as_n+bt_n, is measurable. Now by the monotone convergence theorem and Lemma 10.2.3,

$$\begin{aligned}\int (af+bg)\,d\mu &= \lim_{n\to\infty}\int as_n+bt_n\,d\mu = \lim_{n\to\infty}\left(a\int s_n\,d\mu + b\int t_n\,d\mu\right)\\ &= a\int f\,d\mu + b\int g\,d\mu. \ \blacksquare\end{aligned}$$

As long as you are allowing functions to take the value $+\infty$, you cannot consider something like $f+(-g)$ and so you can't very well expect a satisfactory statement about the integral being linear until you restrict yourself to functions which have values in a vector space. To be linear, a function must be defined on a vector space. This is discussed next.

10.7 The Lebesgue Integral, L^1

The functions considered here have values in $\mathbb{C}$, which is a vector space. A function f with values in $\mathbb{C}$ is of the form $f = \operatorname{Re} f + i\operatorname{Im} f$ where $\operatorname{Re} f$ and $\operatorname{Im} f$ are real valued functions. In fact

$$\operatorname{Re} f = \frac{f+\overline{f}}{2},\ \operatorname{Im} f = \frac{f-\overline{f}}{2i}.$$

Definition 10.7.1 *Let $(\Omega,\mathcal{S},\mu)$ be a measure space and suppose $f:\Omega\to\mathbb{C}$. Then f is said to be measurable if both $\operatorname{Re} f$ and $\operatorname{Im} f$ are measurable real valued functions.*

Of course there is another definition of measurability which says that inverse images of open sets are measurable. This is equivalent to this new definition.

Lemma 10.7.2 *Let $f:\Omega\to\mathbb{C}$. Then f is measurable if and only if $\operatorname{Re} f, \operatorname{Im} f$ are both real valued measurable functions. Also if f,g are complex measurable functions and a,b are complex scalars, then $af+bg$ is also measurable.*

Proof: $\Rightarrow$Suppose first that f is measurable. Recall that $\mathbb{C}$ is considered as $\mathbb{R}^2$ with (x,y) being identified with $x+iy$. Thus the open sets of $\mathbb{C}$ can be obtained with either of the two equivlanent norms $|z| \equiv \sqrt{(\operatorname{Re} z)^2+(\operatorname{Im} z)^2}$ or $\|z\|_\infty = \max(\operatorname{Re} z, \operatorname{Im} z)$. Therefore, if f is measurable

$$\operatorname{Re} f^{-1}(a,b)\cap \operatorname{Im} f^{-1}(c,d) = f^{-1}((a,b)+i(c,d)) \in \mathcal{F}$$

In particular, you could let $(c,d)=\mathbb{R}$ and conclude that $\mathrm{Re}\, f$ is measurable because in this case, the above reduces to the statement that $\mathrm{Re}\, f^{-1}(a,b)\in\mathscr{F}$. Similarly $\mathrm{Im}\, f$ is measurable.

$\Leftarrow$ Next, if each of $\mathrm{Re}\, f$ and $\mathrm{Im}\, f$ are measurable, then

$$f^{-1}((a,b)+i(c,d))=\mathrm{Re}\, f^{-1}(a,b)\cap \mathrm{Im}\, f^{-1}(c,d)\in\mathscr{F}$$

and so, since every open set is the countable union of sets of the form $(a,b)+i(c,d)$, it follows that f is measurable.

Now consider the last claim. Let $h:\mathbb{C}\times\mathbb{C}\to\mathbb{C}$ be given by $h(z,w)\equiv az+bw$. Then h is continuous. If f,g are complex valued measurable functions, consider the complex valued function, $h\circ(f,g):\Omega\to\mathbb{C}$. Then

$$(h\circ(f,g))^{-1}(\text{open})=(f,g)^{-1}\left(h^{-1}(\text{open})\right)=(f,g)^{-1}(\text{open})$$

Now letting U,V be open in $\mathbb{C}$, $(f,g)^{-1}(U\times V)=f^{-1}(U)\cap g^{-1}(V)\in\mathscr{F}$. Since every open set in $\mathbb{C}\times\mathbb{C}$ is the countable union of sets of the form $U\times V$, it follows that $(f,g)^{-1}(\text{open})$ is in $\mathscr{F}$. Thus $af+bg$ is also complex measurable. ■

As is always the case for complex numbers, $|z|^2=(\mathrm{Re}\, z)^2+(\mathrm{Im}\, z)^2$. Also, for g a real valued function, one can consider its positive and negative parts defined respectively as

$$g^+(x)\equiv\frac{g(x)+|g(x)|}{2},\ g^-(x)=\frac{|g(x)|-g(x)}{2}.$$

Thus $|g|=g^++g^-$ and $g=g^+-g^-$ and both g^+ and g^- are measurable nonnegative functions if g is measurable.

Then the following is the definition of what it means for a complex valued function f to be in $L^1(\Omega)$.

Definition 10.7.3 *Let $(\Omega,\mathscr{F},\mu)$ be a measure space. Then a complex valued measurable function f is in $L^1(\Omega)$ if $\int|f|\,d\mu<\infty$. For a function in $L^1(\Omega)$, the integral is defined as follows.*

$$\int f\,d\mu\equiv\int(\mathrm{Re}\, f)^+\,d\mu-\int(\mathrm{Re}\, f)^-\,d\mu+i\left[\int(\mathrm{Im}\, f)^+\,d\mu-\int(\mathrm{Im}\, f)^-\,d\mu\right]$$

I will show that with this definition, the integral is linear and well defined. First note that it is clearly well defined because all the above integrals are of nonnegative functions and are each equal to a nonnegative real number because for h equal to any of the functions, $|h|\le|f|$ and $\int|f|\,d\mu<\infty$.

Here is a lemma which will make it possible to show the integral is linear.

Lemma 10.7.4 *Let g,h,g',h' be nonnegative measurable functions in $L^1(\Omega)$ and suppose that*

$$g-h=g'-h'.$$

Then

$$\int g\,d\mu-\int h\,d\mu=\int g'\,d\mu-\int h'\,d\mu.$$

Proof: By assumption, $g+h'=g'+h$. Then from the Lebesgue integral's righteous algebraic desires, Theorem 10.6.1,

$$\int g d\mu + \int h' d\mu = \int g' d\mu + \int h d\mu$$

which implies the claimed result. ■

Lemma 10.7.5 *Let* $\mathrm{Re}\left(L^1(\Omega)\right)$ *denote the vector space of real valued functions in* $L^1(\Omega)$ *where the field of scalars is the real numbers. Then* $\int d\mu$ *is linear on* $\mathrm{Re}\left(L^1(\Omega)\right)$, *the scalars being real numbers.*

Proof: First observe that from the definition of the positive and negative parts of a function,

$$(f+g)^+ - (f+g)^- = f^+ + g^+ - \left(f^- + g^-\right)$$

because both sides equal $f+g$. Therefore from Lemma 10.7.4 and the definition, it follows from Theorem 10.6.1 that

$$\begin{aligned}\int f+g d\mu &\equiv \int (f+g)^+ - (f+g)^- d\mu = \int f^+ + g^+ d\mu - \int f^- + g^- d\mu \\ &= \int f^+ d\mu + \int g^+ d\mu - \left(\int f^- d\mu + \int g^- d\mu\right) = \int f d\mu + \int g d\mu.\end{aligned}$$

what about taking out scalars? First note that if a is real and nonnegative, then $(af)^+ = af^+$ and $(af)^- = af^-$ while if $a<0$, then $(af)^+ = -af^-$ and $(af)^- = -af^+$. These claims follow immediately from the above definitions of positive and negative parts of a function. Thus if $a<0$ and $f \in L^1(\Omega)$, it follows from Theorem 10.6.1 that

$$\begin{aligned}\int af d\mu &\equiv \int (af)^+ d\mu - \int (af)^- d\mu = \int (-a) f^- d\mu - \int (-a) f^+ d\mu \\ &= -a\int f^- d\mu + a\int f^+ d\mu = a\left(\int f^+ d\mu - \int f^- d\mu\right) \equiv a\int f d\mu.\end{aligned}$$

The case where $a \geq 0$ works out similarly but easier. ■

Now here is the main result.

Theorem 10.7.6 $\int d\mu$ *is linear on* $L^1(\Omega)$ *and* $L^1(\Omega)$ *is a complex vector space. If* $f \in L^1(\Omega)$, *then* $\mathrm{Re}\, f, \mathrm{Im}\, f$, *and* $|f|$ *are all in* $L^1(\Omega)$. *Furthermore, for* $f \in L^1(\Omega)$,

$$\begin{aligned}\int f d\mu &\equiv \int (\mathrm{Re}\, f)^+ d\mu - \int (\mathrm{Re}\, f)^- d\mu + i\left[\int (\mathrm{Im}\, f)^+ d\mu - \int (\mathrm{Im}\, f)^- d\mu\right] \\ &\equiv \int \mathrm{Re}\, f d\mu + i\int \mathrm{Im}\, f d\mu\end{aligned}$$

and the triangle inequality holds,

$$\left|\int f d\mu\right| \leq \int |f| d\mu. \tag{10.4}$$

Also, for every $f \in L^1(\Omega)$ *it follows that for every* $\varepsilon > 0$ *there exists a simple function* s *such that* $|s| \leq |f|$ *and* $\int |f-s| d\mu < \varepsilon$.

Proof: First consider the claim that the integral is linear. It was shown above that the integral is linear on $\mathrm{Re}\left(L^1(\Omega)\right)$. Then letting $a+ib, c+id$ be scalars and f, g functions in $L^1(\Omega)$,

$$(a+ib)f+(c+id)g = (a+ib)(\mathrm{Re}\, f + i\,\mathrm{Im}\, f)+(c+id)(\mathrm{Re}\, g + i\,\mathrm{Im}\, g)$$

$$= c\,\mathrm{Re}(g) - b\,\mathrm{Im}(f) - d\,\mathrm{Im}(g) + a\,\mathrm{Re}(f) + i(b\,\mathrm{Re}(f) + c\,\mathrm{Im}(g) + a\,\mathrm{Im}(f) + d\,\mathrm{Re}(g))$$

It follows from the definition that

$$\int (a+ib)f+(c+id)g\,d\mu = \int (c\,\mathrm{Re}(g) - b\,\mathrm{Im}(f) - d\,\mathrm{Im}(g) + a\,\mathrm{Re}(f))\,d\mu$$

$$+i\int (b\,\mathrm{Re}(f) + c\,\mathrm{Im}(g) + a\,\mathrm{Im}(f) + d\,\mathrm{Re}(g)) \tag{10.5}$$

Also, from the definition,

$$\begin{aligned}(a+ib)\int f\,d\mu + (c+id)\int g\,d\mu &= (a+ib)\left(\int \mathrm{Re}\, f\,d\mu + i\int \mathrm{Im}\, f\,d\mu\right)\\ &\quad +(c+id)\left(\int \mathrm{Re}\, g\,d\mu + i\int \mathrm{Im}\, g\,d\mu\right)\end{aligned}$$

which equals

$$\begin{aligned}&= a\int \mathrm{Re}\, f\,d\mu - b\int \mathrm{Im}\, f\,d\mu + ib\int \mathrm{Re}\, f\,d\mu + ia\int \mathrm{Im}\, f\,d\mu\\ &\quad +c\int \mathrm{Re}\, g\,d\mu - d\int \mathrm{Im}\, g\,d\mu + id\int \mathrm{Re}\, g\,d\mu - d\int \mathrm{Im}\, g\,d\mu.\end{aligned}$$

Using Lemma 10.7.5 and collecting terms, it follows that this reduces to 10.5. Thus the integral is linear as claimed.

Consider the claim about approximation with a simple function. Letting h equal any of

$$(\mathrm{Re}\, f)^+, (\mathrm{Re}\, f)^-, (\mathrm{Im}\, f)^+, (\mathrm{Im}\, f)^-, \tag{10.6}$$

It follows from the monotone convergence theorem and Theorem 9.1.6 on Page 244 there exists a nonnegative simple function $s \le h$ such that

$$\int |h-s|\,d\mu < \frac{\varepsilon}{4}.$$

Therefore, letting s_1, s_2, s_3, s_4 be such simple functions, approximating respectively the functions listed in 10.6, and $s \equiv s_1 - s_2 + i(s_3 - s_4)$,

$$\int |f-s|\,d\mu \le \int \left|(\mathrm{Re}\, f)^+ - s_1\right| d\mu + \int \left|(\mathrm{Re}\, f)^- - s_2\right| d\mu$$

$$+\int \left|(\mathrm{Im}\, f)^+ - s_3\right| d\mu + \int \left|(\mathrm{Im}\, f)^- - s_4\right| d\mu < \varepsilon$$

It is clear from the construction that $|s| \le |f|$.

What about 10.4? Let $\theta \in \mathbb{C}$ be such that $|\theta| = 1$ and $\theta \int f\,d\mu = \left|\int f\,d\mu\right|$. Then from what was shown above about the integral being linear,

$$\left|\int f\,d\mu\right| = \theta\int f\,d\mu = \int \theta f\,d\mu = \int \mathrm{Re}(\theta f)\,d\mu \le \int |f|\,d\mu.$$

If $f,g \in L^1(\Omega)$, then it is known that for a,b scalars, it follows that $af+bg$ is measurable. See Lemma 10.7.2. Also

$$\int |af+bg|\,d\mu \leq \int |a|\,|f|+|b|\,|g|\,d\mu < \infty. \blacksquare$$

The following corollary follows from this. The conditions of this corollary are sometimes taken as a definition of what it means for a function f to be in $L^1(\Omega)$.

Corollary 10.7.7 *$f \in L^1(\Omega)$ if and only if there exists a sequence of complex simple functions, $\{s_n\}$ such that*

$$\begin{array}{l} s_n(\omega) \to f(\omega) \text{ for all } \omega \in \Omega \\ \lim_{m,n\to\infty} \int (|s_n - s_m|)\,d\mu = 0 \end{array} \tag{10.7}$$

When $f \in L^1(\Omega)$,

$$\int f d\mu \equiv \lim_{n\to\infty} \int s_n. \tag{10.8}$$

Proof: From the above theorem, if $f \in L^1$ there exists a sequence of simple functions $\{s_n\}$ such that

$$\int |f-s_n|\,d\mu < 1/n,\ s_n(\omega) \to f(\omega) \text{ for all } \omega$$

Then

$$\int |s_n - s_m|\,d\mu \leq \int |s_n - f|\,d\mu + \int |f-s_m|\,d\mu \leq \frac{1}{n} + \frac{1}{m}.$$

Next suppose the existence of the approximating sequence of simple functions. Then f is measurable because its real and imaginary parts are the limit of measurable functions. By Fatou's lemma, $\int |f|\,d\mu \leq \liminf_{n\to\infty} \int |s_n|\,d\mu < \infty$ because $|\int |s_n|\,d\mu - \int |s_m|\,d\mu| \leq \int |s_n - s_m|\,d\mu$ which is given to converge to 0. Thus $\{\int |s_n|\,d\mu\}$ is a Cauchy sequence and is therefore, bounded.

In case $f \in L^1(\Omega)$, letting $\{s_n\}$ be the approximating sequence, Fatou's lemma implies

$$\left| \int f d\mu - \int s_n d\mu \right| \leq \int |f-s_n|\,d\mu \leq \lim \inf_{m\to\infty} \int |s_m - s_n|\,d\mu < \varepsilon$$

provided n is large enough. Hence 10.8 follows. $\blacksquare$

This is a good time to observe the following fundamental observation which follows from a repeat of the above arguments.

Theorem 10.7.8 *Suppose $\Lambda(f) \in [0,\infty]$ for all nonnegative measurable functions and suppose that for $a,b \geq 0$ and f,g nonnegative measurable functions,*

$$\Lambda(af+bg) = a\Lambda(f) + b\Lambda(g).$$

In other words, Λ wants to be linear. Then Λ has a unique linear extension to the set of measurable functions

$$\{f \text{ measurable } : \Lambda(|f|) < \infty\},$$

this set being a vector space.

10.8 The Dominated Convergence Theorem

One of the major theorems in this theory is the dominated convergence theorem. Before presenting it, here is a technical lemma about lim sup and lim inf which is really pretty obvious from the definition.

Lemma 10.8.1 *Let $\{a_n\}$ be a sequence in $[-\infty,\infty]$. Then $\lim_{n\to\infty} a_n$ exists if and only if $\liminf_{n\to\infty} a_n = \limsup_{n\to\infty} a_n$ and in this case, the limit equals the common value of these two numbers.*

Proof: Suppose first $\lim_{n\to\infty} a_n = a \in \mathbb{R}$. Letting $\varepsilon > 0$ be given, $a_n \in (a-\varepsilon, a+\varepsilon)$ for all n large enough, say $n \geq N$. Therefore, both $\inf\{a_k : k \geq n\}$ and $\sup\{a_k : k \geq n\}$ are contained in $[a-\varepsilon, a+\varepsilon]$ whenever $n \geq N$. It follows $\limsup_{n\to\infty} a_n$ and $\liminf_{n\to\infty} a_n$ are both in $[a-\varepsilon, a+\varepsilon]$, showing

$$\left| \liminf_{n\to\infty} a_n - \limsup_{n\to\infty} a_n \right| < 2\varepsilon.$$

Since ε is arbitrary, the two must be equal and they both must equal a. Next suppose $\lim_{n\to\infty} a_n = \infty$. Then if $l \in \mathbb{R}$, there exists N such that for $n \geq N, l \leq a_n$ and therefore, for such n,

$$l \leq \inf\{a_k : k \geq n\} \leq \sup\{a_k : k \geq n\}$$

and this shows, since l is arbitrary that $\liminf_{n\to\infty} a_n = \limsup_{n\to\infty} a_n = \infty$. The case for $-\infty$ is similar.

Conversely, suppose $\liminf_{n\to\infty} a_n = \limsup_{n\to\infty} a_n = a$. Suppose first that $a \in \mathbb{R}$. Then, letting $\varepsilon > 0$ be given, there exists N such that if $n \geq N$,

$$\sup\{a_k : k \geq n\} - \inf\{a_k : k \geq n\} < \varepsilon$$

therefore, if $k, m > N$, and $a_k > a_m$,

$$|a_k - a_m| = a_k - a_m \leq \sup\{a_k : k \geq n\} - \inf\{a_k : k \geq n\} < \varepsilon$$

showing that $\{a_n\}$ is a Cauchy sequence. Therefore, it converges to $a \in \mathbb{R}$, and as in the first part, the lim inf and lim sup both equal a. If $\liminf_{n\to\infty} a_n = \limsup_{n\to\infty} a_n = \infty$, then given $l \in \mathbb{R}$, there exists N such that for $n \geq N$, $\inf_{n>N} a_n > l$.Therefore,

$$\lim_{n\to\infty} a_n = \infty.$$

The case for $-\infty$ is similar. ■

Here is the dominated convergence theorem.

Theorem 10.8.2 *(Dominated Convergence theorem) Let $f_n \in L^1(\Omega)$ and suppose*

$$f(\omega) = \lim_{n\to\infty} f_n(\omega),$$

and there exists a measurable function g, with values in $[0,\infty]$,[1] such that

$$|f_n(\omega)| \leq g(\omega) \text{ and } \int g(\omega)\, d\mu < \infty.$$

[1] Note that, since g is allowed to have the value ∞, it is not known that $g \in L^1(\Omega)$.

Then $f \in L^1(\Omega)$ *and*

$$0 = \lim_{n\to\infty} \int |f_n - f| \, d\mu = \lim_{n\to\infty} \left| \int f d\mu - \int f_n d\mu \right|$$

Proof: f is measurable by Theorem 9.1.2. Since $|f| \leq g$, it follows that

$$f \in L^1(\Omega) \text{ and } |f - f_n| \leq 2g.$$

By Fatou's lemma (Theorem 10.5.1),

$$\int 2g d\mu \leq \liminf_{n\to\infty} \int 2g - |f - f_n| d\mu = \int 2g d\mu - \limsup_{n\to\infty} \int |f - f_n| d\mu.$$

Subtracting $\int 2g d\mu$, $0 \leq -\limsup_{n\to\infty} \int |f - f_n| d\mu$. Hence

$$\begin{aligned} 0 &\geq \limsup_{n\to\infty} \left(\int |f - f_n| d\mu \right) \\ &\geq \liminf_{n\to\infty} \left(\int |f - f_n| d\mu \right) \geq \liminf_{n\to\infty} \left| \int f d\mu - \int f_n d\mu \right| \geq 0. \end{aligned}$$

This proves the theorem by Lemma 10.8.1 because the lim sup and lim inf are equal. ■

Corollary 10.8.3 *Suppose* $f_n \in L^1(\Omega)$ *and* $f(\omega) = \lim_{n\to\infty} f_n(\omega)$. *Suppose also there exist measurable functions,* g_n, g *with values in* $[0,\infty]$ *such that*

$$\lim_{n\to\infty} \int g_n d\mu = \int g d\mu$$

$g_n(\omega) \to g(\omega)$ μ *a.e. and both* $\int g_n d\mu$ *and* $\int g d\mu$ *are finite. Also suppose* $|f_n(\omega)| \leq g_n(\omega)$. *Then* $\lim_{n\to\infty} \int |f - f_n| d\mu = 0$.

Proof: It is just like the above. This time $g + g_n - |f - f_n| \geq 0$ and so by Fatou's lemma,

$$\int 2g d\mu - \limsup_{n\to\infty} \int |f - f_n| d\mu = \lim_{n\to\infty} \int (g_n + g) d\mu - \limsup_{n\to\infty} \int |f - f_n| d\mu$$

$$\begin{aligned} &= \liminf_{n\to\infty} \int (g_n + g) d\mu - \limsup_{n\to\infty} \int |f - f_n| d\mu \\ &= \liminf_{n\to\infty} \int ((g_n + g) - |f - f_n|) d\mu \geq \int 2g d\mu \end{aligned}$$

and so $-\limsup_{n\to\infty} \int |f - f_n| d\mu \geq 0$. Thus

$$\begin{aligned} 0 &\geq \limsup_{n\to\infty} \left(\int |f - f_n| d\mu \right) \\ &\geq \liminf_{n\to\infty} \left(\int |f - f_n| d\mu \right) \geq \left| \int f d\mu - \int f_n d\mu \right| \geq 0. \blacksquare \end{aligned}$$

Definition 10.8.4 *Let* E *be a measurable subset of* Ω.

$$\int_E f d\mu \equiv \int f \mathcal{X}_E d\mu.$$

If $L^1(E)$ is written, the σ algebra is defined as $\{E \cap A : A \in \mathscr{F}\}$ and the measure is μ restricted to this smaller σ algebra. Clearly, if $f \in L^1(\Omega)$, then $f\mathscr{X}_E \in L^1(E)$ and if $f \in L^1(E)$, then letting $\tilde{f}$ be the 0 extension of f off of E, it follows $\tilde{f} \in L^1(\Omega)$.

Another very important observation applies to the case where Ω is also a metric space. In this lemma, $\text{spt}(f)$ denotes the closure of the set on which f is nonzero.

Definition 10.8.5 *Let K be a set and let V be an open set containing K. Then the notation $K \prec f \prec V$ means that $f(x) = 1$ for all $x \in K$ and $\text{spt}(f)$ is a compact subset of V. $\text{spt}(f)$ is defined as the closure of the set where f is not zero. It is called the "support" of f. A function $f \in C_c(\Omega)$ for Ω a metric space if f is continuous on Ω and $\text{spt}(f)$ is compact. This $C_c(\Omega)$ is called the continuous functions with compact support.*

Recall Lemma 3.12.4. Listed next for convenience.

Lemma 10.8.6 *Let Ω be a metric space in which the closed balls are compact and let K be a compact subset of V, an open set. Then there exists a continuous function $f : \Omega \to [0,1]$ such that $K \prec f \prec V$.*

Theorem 10.8.7 *Let $(\Omega, \mathscr{S}, \mu)$ be a regular measure space, meaning that μ is inner and outer regular and $\mu(K) < \infty$ for each compact set K. Suppose also that the conclusion of Lemma 3.12.4 holds. Then for each $\varepsilon > 0$ and $f \in L^1(\Omega)$, there is $g \in C_c(\Omega)$ such that $\int_\Omega |f - g| \, d\mu < \varepsilon$.*

Proof: First consider a measurable set E where $\mu(E) < \infty$. Let $K \subseteq E \subseteq V$ where $\mu(V \setminus K) < \varepsilon$. Now let $K \prec h \prec V$. Then

$$\int |h - \mathscr{X}_E| \, d\mu \leq \int \mathscr{X}_{V \setminus K} d\mu = \mu(V \setminus K) < \varepsilon. \tag{10.9}$$

By Corollary 10.7.7, there is a sequence of simple functions converging pointwise to f such that for $m, n > N$, $\frac{\varepsilon}{2} > \int (|s_n - s_m|) \, d\mu$. Then let $n \to \infty$ and apply the dominated convergence theorem to get $\int |f - s_m| \, d\mu \leq \frac{\varepsilon}{2}$. However, from 10.9, there is g in $C_c(\Omega)$ such that $\int |s_m - g| \, d\mu < \frac{\varepsilon}{2}$ and so

$$\int |f - g| \, d\mu \leq \int |f - s_m| \, d\mu + \int |s_m - g| \, d\mu < \varepsilon. \blacksquare$$

10.9 Some Important General Theory

10.9.1 Eggoroff's Theorem

You might show that a sequence of measurable real or complex valued functions converges on a measurable set. This is Proposition 9.1.8 above. Eggoroff's theorem says that if the set of points where a sequence of measurable functions converges is all but a set of measure zero, then the sequence almost converges uniformly in a certain sense.

Theorem 10.9.1 *(Egoroff) Let $(\Omega, \mathscr{F}, \mu)$ be a finite measure space, $\mu(\Omega) < \infty$ and let f_n, f be complex valued functions such that $\text{Re}\, f_n, \text{Im}\, f_n$ are all measurable and*

$$\lim_{n \to \infty} f_n(\omega) = f(\omega)$$

for all $\omega \notin E$ where $\mu(E) = 0$. Then for every $\varepsilon > 0$, there exists a set,

$$F \supseteq E,\ \mu(F) < \varepsilon,$$

such that f_n converges uniformly to f on F^C.

Proof: First suppose $E = \emptyset$ so that convergence is pointwise everywhere. It follows then that $\mathrm{Re}\, f$ and $\mathrm{Im}\, f$ are pointwise limits of measurable functions and are therefore measurable. Let

$$E_{km} = \{\omega \in \Omega : |f_n(\omega) - f(\omega)| \geq 1/m \text{ for some } n > k\}.$$

Note that

$$|f_n(\omega) - f(\omega)| = \sqrt{(\mathrm{Re}\, f_n(\omega) - \mathrm{Re}\, f(\omega))^2 + (\mathrm{Im}\, f_n(\omega) - \mathrm{Im}\, f(\omega))^2}$$

and so, $\left[|f_n - f| \geq \frac{1}{m}\right]$ is measurable. Hence E_{km} is measurable because

$$E_{km} = \cup_{n=k+1}^{\infty} \left[|f_n - f| \geq \frac{1}{m}\right].$$

For fixed $m, \cap_{k=1}^{\infty} E_{km} = \emptyset$ because f_n converges to f. Therefore, if $\omega \in \Omega$ there exists k such that if $n > k$, $|f_n(\omega) - f(\omega)| < \frac{1}{m}$ which means $\omega \notin E_{km}$. Note also that $E_{km} \supseteq E_{(k+1)m}$. Since $\mu(E_{1m}) < \infty$, Theorem 9.2.4 on Page 246 implies

$$0 = \mu(\cap_{k=1}^{\infty} E_{km}) = \lim_{k \to \infty} \mu(E_{km}).$$

Let $k(m)$ be chosen such that $\mu(E_{k(m)m}) < \varepsilon 2^{-m}$ and let $F = \cup_{m=1}^{\infty} E_{k(m)m}$. Then $\mu(F) < \varepsilon$ because $\mu(F) \leq \sum_{m=1}^{\infty} \mu\left(E_{k(m)m}\right) < \sum_{m=1}^{\infty} \varepsilon 2^{-m} = \varepsilon$.

Now let $\eta > 0$ be given and pick m_0 such that $m_0^{-1} < \eta$. If $\omega \in F^C$, then $\omega \in \bigcap_{m=1}^{\infty} E_{k(m)m}^C$. Hence $\omega \in E_{k(m_0)m_0}^C$ so $|f_n(\omega) - f(\omega)| < 1/m_0 < \eta$ for all $n > k(m_0)$. This holds for all $\omega \in F^C$ and so f_n converges uniformly to f on F^C.

Now if $E \neq \emptyset$, consider $\{\mathscr{X}_{E^C} f_n\}_{n=1}^{\infty}$. Each $\mathscr{X}_{E^C} f_n$ has real and imaginary parts measurable and the sequence converges pointwise to $\mathscr{X}_{E^C} f$ everywhere. Therefore, from the first part, there exists a set of measure less than ε, F such that on $F^C, \{\mathscr{X}_{E^C} f_n\}$ converges uniformly to $\mathscr{X}_{E^C} f$. Therefore, on $(E \cup F)^C$, $\{f_n\}$ converges uniformly to f. This proves the theorem. ■

10.9.2 The Vitali Convergence Theorem

The Vitali convergence theorem is a convergence theorem which in the case of a finite measure space is superior to the dominated convergence theorem.

Definition 10.9.2 *Let $(\Omega, \mathscr{F}, \mu)$ be a measure space and let $\mathfrak{S} \subseteq L^1(\Omega)$. $\mathfrak{S}$ is uniformly integrable if for every $\varepsilon > 0$ there exists $\delta > 0$ such that for all $f \in \mathfrak{S}$*

$$\left|\int_E f d\mu\right| < \varepsilon \text{ whenever } \mu(E) < \delta.$$

Lemma 10.9.3 *If $\mathcal{S}$ is uniformly integrable, then $|\mathcal{S}| \equiv \{|f| : f \in \mathcal{S}\}$ is uniformly integrable. Also $\mathcal{S}$ is uniformly integrable if $\mathcal{S}$ is finite.*

Proof: Let $\varepsilon > 0$ be given and suppose $\mathcal{S}$ is uniformly integrable. First suppose the functions are real valued. Let δ be such that if $\mu(E) < \delta$, then $|\int_E f d\mu| < \frac{\varepsilon}{2}$ for all $f \in \mathcal{S}$. Let $\mu(E) < \delta$. Then if $f \in \mathcal{S}$,

$$\begin{aligned}\int_E |f|\, d\mu &\leq \int_{E\cap[f\leq 0]} (-f)\, d\mu + \int_{E\cap[f>0]} f d\mu = \left|\int_{E\cap[f\leq 0]} f d\mu\right| + \left|\int_{E\cap[f>0]} f d\mu\right| \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.\end{aligned}$$

In general, if $\mathcal{S}$ is a uniformly integrable set of complex valued functions, the inequalities,

$$\left|\int_E \mathrm{Re}\, f d\mu\right| \leq \left|\int_E f d\mu\right|, \left|\int_E \mathrm{Im}\, f d\mu\right| \leq \left|\int_E f d\mu\right|,$$

imply $\mathrm{Re}\,\mathcal{S} \equiv \{\mathrm{Re}\, f : f \in \mathcal{S}\}$ and $\mathrm{Im}\,\mathcal{S} \equiv \{\mathrm{Im}\, f : f \in \mathcal{S}\}$ are also uniformly integrable. Therefore, applying the above result for real valued functions to these sets of functions, it follows $|\mathcal{S}|$ is uniformly integrable also.

For the last part, is suffices to verify a single function in $L^1(\Omega)$ is uniformly integrable. To do so, note that from the dominated convergence theorem,

$$\lim_{R\to\infty} \int_{[|f|>R]} |f|\, d\mu = 0.$$

Let $\varepsilon > 0$ be given and choose R large enough that $\int_{[|f|>R]} |f|\, d\mu < \frac{\varepsilon}{2}$. Now let $\mu(E) < \frac{\varepsilon}{2R}$. Then

$$\begin{aligned}\int_E |f|\, d\mu &= \int_{E\cap[|f|\leq R]} |f|\, d\mu + \int_{E\cap[|f|>R]} |f|\, d\mu \\ &< R\mu(E) + \frac{\varepsilon}{2} < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.\end{aligned}$$

This proves the lemma. ■

The following gives a nice way to identify a uniformly integrable set of functions.

Lemma 10.9.4 *Let $\mathcal{S}$ be a subset of $L^1(\Omega,\mu)$ where $\mu(\Omega) < \infty$. Let*

$$t \to h(t)$$

be a continuous function which satisfies

$$\lim_{t\to\infty} \frac{h(t)}{t} = \infty$$

Then $\mathcal{S}$ is uniformly integrable and bounded in $L^1(\Omega)$ if

$$\sup\left\{\int_\Omega h(|f|)\, d\mu : f \in \mathcal{S}\right\} = N < \infty.$$

Proof: First I show $\mathcal{S}$ is bounded in $L^1(\Omega;\mu)$ which means there exists a constant M such that for all $f \in \mathcal{S}$, $\int_\Omega |f|\,d\mu \le M$. From the properties of h, there exists R_n such that if $t \ge R_n$, then $h(t) \ge nt$. Therefore,

$$\int_\Omega |f|\,d\mu = \int_{[|f|\ge R_n]} |f|\,d\mu + \int_{[|f|<R_n]} |f|\,d\mu$$

Letting $n = 1$, and $f \in \mathcal{S}$,

$$\begin{aligned}\int_\Omega |f|\,d\mu &= \int_{[|f|\ge R_1]} |f|\,d\mu + \int_{[|f|<R_1]} |f|\,d\mu \\ &\le \int_{[|f|\ge R_1]} h(|f|)\,d\mu + R_1\mu([|f| < R_1]) \\ &\le N + R_1\mu(\Omega) \equiv M. \end{aligned} \tag{10.10}$$

Next let E be a measurable set. Then for every $f \in \mathcal{S}$, it follows from 10.10

$$\begin{aligned}\int_E |f|\,d\mu &= \int_{[|f|\ge R_n]\cap E} |f|\,d\mu + \int_{[|f|<R_n]\cap E} |f|\,d\mu \\ &\le \frac{1}{n}\int_\Omega |f|\,d\mu + R_n\mu(E) \le \frac{M}{n} + R_n\mu(E) \end{aligned} \tag{10.11}$$

Let n be large enough that $M/n < \varepsilon/2$ and then let $\mu(E) < \varepsilon/2R_n$. Then 10.11 is less than $\varepsilon/2 + R_n(\varepsilon/2R_n) = \varepsilon$ ■

Letting $h(t) = t^2$, it follows that if all the functions in $\mathcal{S}$ are bounded, then the collection of functions is uniformly integrable. Another way to discuss uniform integrability is the following. This other way involving equi-integrability is used a lot in probability.

Definition 10.9.5 *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space with $\mu(\Omega) < \infty$. A set $\mathcal{S} \subseteq L^1(\Omega)$ is said to be equi-integrable if for every $\varepsilon > 0$ there exists $\lambda > 0$ such that*

$$\int_{[|f|>\lambda]} |f|\,d\mu < \varepsilon$$

for all $f \in \mathcal{S}$.

Then the relation between this and uniform integrability is as follows.

Proposition 10.9.6 *In the context of the above definition, $\mathcal{S}$ is equi-integrable if and only if it is a bounded subset of $L^1(\Omega)$ which is also uniformly integrable.*

Proof: $\Rightarrow$ I need to show $\mathcal{S}$ is bounded and uniformly integrable. First consider bounded. Choose λ to work for $\varepsilon = 1$. Then for all $f \in \mathcal{S}$,

$$\int |f|\,d\mu = \int_{[|f|>\lambda]} |f|\,d\mu + \int_{[|f|\le\lambda]} |f| \le 1 + \lambda\mu(\Omega)$$

Thus it is bounded. Now let E be a measurable subset of Ω. Let λ go with $\varepsilon/2$ in the definition of equi-integrable. Then for all $f \in \mathcal{S}$,

$$\int_E |f|\,d\mu \le \int_{[|f|>\lambda]} |f|\,d\mu + \int_{E\cap[|f|\le\lambda]} |f|\,d\mu \le \frac{\varepsilon}{2} + \lambda\mu(E)$$

Then let $\mu(E)$ be small enough that $\lambda\mu(E) < \varepsilon/2$ and this shows uniform integrability.

$\Leftarrow$ I need to verify equi-integrable from bounded and uniformly integrable. Let δ be such that if $\mu(E) < \delta$, then $\int_E |f|\,d\mu < \varepsilon$ for all $f \in \mathfrak{S}$. If not, then there exists $f_n \in \mathfrak{S}$ with $[|f_n| > n] > \delta$. Thus $\int |f_n|\,d\mu \geq \int_{[|f_n|>n]} |f_n|\,d\mu \geq n\mu([|f_n| > n]) > n\delta$ and so $\mathfrak{S}$ is not bounded after all. ■

The following theorem is Vitali's convergence theorem.

Theorem 10.9.7 *Let $\{f_n\}$ be a uniformly integrable set of complex valued functions, $\mu(\Omega) < \infty$, and $f_n(x) \to f(x)$ a.e. where f is a measurable complex valued function. Then $f \in L^1(\Omega)$ and*

$$\lim_{n\to\infty} \int_\Omega |f_n - f|\,d\mu = 0. \tag{10.12}$$

Proof: First it will be shown that $f \in L^1(\Omega)$. By uniform integrability, there exists $\delta > 0$ such that if $\mu(E) < \delta$, then $\int_E |f_n|\,d\mu < 1$ for all n. By Egoroff's theorem, there exists a set E of measure less than δ such that on E^C, $\{f_n\}$ converges uniformly. Therefore, for p large enough, and $n > p$, $\int_{E^C} |f_p - f_n|\,d\mu < 1$ which implies $\int_{E^C} |f_n|\,d\mu < 1 + \int_\Omega |f_p|\,d\mu$.Then since there are only finitely many functions, f_n with $n \leq p$, there exists a constant, M_1 such that for all n, $\int_{E^C} |f_n|\,d\mu < M_1$. But also,

$$\int_\Omega |f_m|\,d\mu = \int_{E^C} |f_m|\,d\mu + \int_E |f_m| \leq M_1 + 1 \equiv M.$$

Therefore, by Fatou's lemma, $\int_\Omega |f|\,d\mu \leq \liminf_{n\to\infty} \int |f_n|\,d\mu \leq M$, showing that $f \in L^1$ as hoped.

Now $\mathfrak{S} \cup \{f\}$ is uniformly integrable so there exists $\delta_1 > 0$ such that if $\mu(E) < \delta_1$, then $\int_E |g|\,d\mu < \varepsilon/3$ for all $g \in \mathfrak{S} \cup \{f\}$.

By Egoroff's theorem, there exists a set, F with $\mu(F) < \delta_1$ such that f_n converges uniformly to f on F^C. Therefore, there exists m such that if $n > m$, then $\int_{F^C} |f - f_n|\,d\mu < \frac{\varepsilon}{3}$. It follows that for $n > m$,

$$\begin{aligned} \int_\Omega |f - f_n|\,d\mu &\leq \int_{F^C} |f - f_n|\,d\mu + \int_F |f|\,d\mu + \int_F |f_n|\,d\mu \\ &< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon, \end{aligned}$$

which verifies 10.12. ■

10.10 One Dimensional Lebesgue Stieltjes Integral

Let F be an increasing function defined on $\mathbb{R}$. Let μ be the Lebesgue Stieltjes measure defined in Theorems 9.10.1 and 9.7.4. The conclusions of these theorems are reviewed here.

Theorem 10.10.1 *Let F be an increasing function defined on $\mathbb{R}$, an integrator function. There exists a function $\mu : \mathscr{P}(\mathbb{R}) \to [0,\infty]$ which satisfies the conditions of Theorem 9.7.4 in terms of measures of intervals and the inner and outer regularity properties.*

The Lebesgue integral taken with respect to this measure, is called the Lebesgue Stieltjes integral. Note that any real valued continuous function is measurable with respect to $\mathscr{S}$. This is because if f is continuous, inverse images of open sets are open and open sets are in

$\mathscr{S}$. Thus f is measurable because $f^{-1}((a,b)) \in \mathscr{S}$. Similarly if f has complex values this argument applied to its real and imaginary parts yields the conclusion that f is measurable. This will be denoted here by $\int f d\mu$ but it is often the case that it is denoted as $\int f dF$.

In the case of most interest, where $F(x) = x$, how does the Lebesgue integral compare with the Riemann integral? The short answer is that if f is Riemann integrable, then it is also Lebesgue interable and the two integrals coincide. It is customary to denote the Lebesgue integral in this context as $\int_a^b f dm$.

Theorem 10.10.2 *Suppose f is Riemann integrable on an interval $[a,b]$. Then f is also Lebesgue integrable and the two integrals are the same.*

Proof: It suffices to consider the case that f is nonnegative. Otherwise, one simply considers the positive and negative parts of the real and imaginary parts of the function. Thus f is a bounded function and there is a decreasing sequence of upper step functions, denoted as $\{u_n\}$ and an increasing sequence of lower step functions denoted as $\{l_n\}$ such that

$$\int_a^b l_n dt \leq \int_a^b f dt \leq \int_a^b u_n dt, \left| \int_a^b u_n dt - \int_a^b l_n dt \right| < 2^{-n}$$

Since f must be bounded, it can be assumed that $|u_n(t)|, |l_n(t)| < M$ for some constant M. Let $g(t) = \lim_{n\to\infty} u_n(t)$ and $h(t) = \lim_{n\to\infty} l_n(t)$. Then from the dominated convergence theorem (Why?) one obtains

$$\begin{aligned} \int_a^b f dt &= \lim_{n\to\infty} \int_a^b l_n dt = \lim_{n\to\infty} \int_a^b l_n dm = \int_a^b h dm \\ &\leq \int_a^b g dm \leq \lim_{n\to\infty} \int_a^b u_n dm = \lim_{n\to\infty} \int_a^b u_n dt = \int_a^b f dt \end{aligned}$$

Also, from the construction, $h(t) \leq f(t) \leq g(t)$. From the above,

$$\int_a^b |g(t) - h(t)| dm = 0.$$

It follows that g is measurable (why?) and $f(t) = g(t)$ for m a.e. t. (why?) By completeness of the measure, it follows that f is Lebesgue measurable and $\int_a^b f dm = \int_a^b g dm = \int_a^b h dm = \int_a^b f dt$. (why?) ■

If you have seen the Darboux Stieltjes integral, defined like the Riemann integral in terms of upper and lower sums, the following compares the Lebesgue Stieltjes integral with this one also. For f a continuous function, how does the Lebesgue Stieltjes integral compare with the Darboux Stieltjes integral? To answer this question, here is a technical lemma.

Lemma 10.10.3 *Let D be a countable subset of $\mathbb{R}$ and suppose $a,b \notin D$. Also suppose f is a continuous function defined on $[a,b]$. Then there exists a sequence of functions $\{s_n\}$ of the form*

$$s_n(x) \equiv \sum_{k=1}^{m_n} f\left(z_{k-1}^n\right) \mathscr{X}_{[z_{k-1}^n, z_k^n)}(x)$$

such that each $z_k^n \notin D$ and

$$\sup\{|s_n(x) - f(x)| : x \in [a,b]\} < 1/n.$$

Proof: First note that D contains no intervals. To see this let $D=\{d_k\}_{k=1}^{\infty}$. If D has an interval of length 2ε, let I_k be an interval centered at d_k which has length $\varepsilon/2^k$. Therefore, the sum of the lengths of these intervals is no more than $\sum_{k=1}^{\infty}\frac{\varepsilon}{2^k}=\varepsilon$. Thus D cannot contain an interval of length 2ε. Since ε is arbitrary, D cannot contain any interval.

Since f is continuous, it follows from Theorem 3.7.4 on Page 82 that f is uniformly continuous. Therefore, there exists $\delta>0$ such that if $|x-y|\leq 3\delta$, then $|f(x)-f(y)|<1/n$. Now let $\{x_0,\cdots,x_{m_n}\}$ be a partition of $[a,b]$ such that $|x_i-x_{i-1}|<\delta$ for each i. For $k=1,2,\cdots,m_n-1$, let $z_k^n\notin D$ and $\left|z_k^n-x_k\right|<\delta$. Then

$$\left|z_k^n-z_{k-1}^n\right|\leq\left|z_k^n-x_k\right|+\left|x_k-x_{k-1}\right|+\left|x_{k-1}-z_{k-1}^n\right|<3\delta.$$

It follows that for each $x\in[a,b]$, $\left|\sum_{k=1}^{m_n}f\left(z_{k-1}^n\right)\mathcal{X}_{[z_{k-1}^n,z_k^n)}(x)-f(x)\right|<1/n$. ■

Proposition 10.10.4 *Let f be a continuous function defined on $\mathbb{R}$. Also let F be an increasing function defined on $\mathbb{R}$. Then whenever c,d are not points of discontinuity of F and $[a,b]\supseteq[c,d]$, $\int_a^b f\mathcal{X}_{[c,d]}dF=\int f\mathcal{X}_{[c,d]}d\mu$.Here μ is the Lebesgue Stieltjes measure defined above.*

Proof: Since F is an increasing function it can have only countably many discontinuities. The reason for this is that the only kind of discontinuity it can have is where $F(x+)>F(x-)$. Now since F is increasing, the intervals $(F(x-),F(x+))$ for x a point of discontinuity are disjoint and so since each must contain a rational number and the rational numbers are countable, and therefore so are these intervals.

Let D denote this countable set of discontinuities of F. Then if $l,r\notin D,[l,r]\subseteq[a,b]$, it follows quickly from the definition of the Darboux Stieltjes integral that

$$\int_a^b\mathcal{X}_{[l,r)}dF=F(r)-F(l)=F(r-)-F(l-)=\mu([l,r))=\int\mathcal{X}_{[l,r)}d\mu.$$

Now let $\{s_n\}$ be the sequence of step functions of Lemma 10.10.3 such that these step functions converge uniformly to f on $[c,d]$, say $\max_x|f(x)-s_n(x)|<1/n$. Then

$$\left|\int\left(\mathcal{X}_{[c,d]}f-\mathcal{X}_{[c,d]}s_n\right)d\mu\right|\leq\int\left|\mathcal{X}_{[c,d]}(f-s_n)\right|d\mu\leq\frac{1}{n}\mu([c,d])$$

and

$$\left|\int_a^b\left(\mathcal{X}_{[c,d]}f-\mathcal{X}_{[c,d]}s_n\right)dF\right|\leq\int_a^b\mathcal{X}_{[c,d]}|f-s_n|dF<\frac{1}{n}(F(b)-F(a)).$$

Also if s_n is given by the formula of Lemma 10.10.3,

$$\int\mathcal{X}_{[c,d]}s_nd\mu=\int\sum_{k=1}^{m_n}f\left(z_{k-1}^n\right)\mathcal{X}_{[z_{k-1}^n,z_k^n)}d\mu=\sum_{k=1}^{m_n}\int f\left(z_{k-1}^n\right)\mathcal{X}_{[z_{k-1}^n,z_k^n)}d\mu$$

$$=\sum_{k=1}^{m_n}f\left(z_{k-1}^n\right)\mu\left([z_{k-1}^n,z_k^n)\right)=\sum_{k=1}^{m_n}f\left(z_{k-1}^n\right)\left(F\left(z_k^n-\right)-F\left(z_{k-1}^n-\right)\right)$$

$$=\sum_{k=1}^{m_n}f\left(z_{k-1}^n\right)\left(F\left(z_k^n\right)-F\left(z_{k-1}^n\right)\right)=\sum_{k=1}^{m_n}\int_a^b f\left(z_{k-1}^n\right)\mathcal{X}_{[z_{k-1}^n,z_k^n)}dF=\int_a^b s_ndF.$$

Therefore,

$$\left|\int \mathcal{X}_{[c,d]} f d\mu - \int_a^b \mathcal{X}_{[c,d]} f dF\right| \leq \left|\int \mathcal{X}_{[c,d]} f d\mu - \int \mathcal{X}_{[c,d]} s_n d\mu\right|$$

$$+\left|\int \mathcal{X}_{[c,d]} s_n d\mu - \int_a^b s_n dF\right| + \left|\int_a^b s_n dF - \int_a^b \mathcal{X}_{[c,d]} f dF\right|$$

$$\leq \frac{1}{n}\mu([c,d]) + \frac{1}{n}(F(b) - F(a))$$

and since n is arbitrary, this shows

$$\int f d\mu - \int_a^b f dF = 0. \blacksquare$$

In particular, in the special case where F is continuous and f is continuous, $\int_a^b f dF = \int \mathcal{X}_{[a,b]} f d\mu$. Thus, if $F(x) = x$ so the Darboux Stieltjes integral is the usual integral from calculus, $\int_a^b f(t)\,dt = \int \mathcal{X}_{[a,b]} f d\mu$ where μ is the measure which comes from $F(x) = x$ as described above. This measure is often denoted by m. Thus when f is continuous $\int_a^b f(t)\,dt = \int \mathcal{X}_{[a,b]} f dm$ and so there is no problem in writing $\int_a^b f(t)\,dt$ for either the Lebesgue or the Riemann integral. Furthermore, when f is continuous, you can compute the Lebesgue integral by using the fundamental theorem of calculus because in this case, the two integrals are equal. $\blacksquare$

Note that as a special case, if f is continuous on $\mathbb{R}$ and equals 0 off some finite interval I, written as $f \in C_c(\mathbb{R})$, then $\int_{\mathbb{R}} f dF = \int_{\mathbb{R}} f d\mu$ where the first integral is defined as $\int_c^d f dF$ where $(c,d) \supseteq I$. You could use the Riemann Stieltjes integral to define a positive linear functional on $C_c(\mathbb{R})$ as just explained and then it follows that the Lebesuge integral taken with respect to the Lebesgue Stieltjes measure above equals this functional on $C_c(\mathbb{R})$. This idea will be discussed more later in the abstract theory when such functionals will be shown to determine measures.

10.11 The Distribution Function

For $(\Omega, \mathcal{F}, \mu)$ a measure space, the integral of a nonnegative measurable function was defined earlier as

$$\int f d\mu \equiv \int_0^\infty \mu([f > t])\,dt$$

This idea will be developed more in this section.

Definition 10.11.1 *Let $f \geq 0$ and suppose f is measurable. The distribution function is the function defined by $t \to \mu([t < f])$.*

Lemma 10.11.2 *If $\{f_n\}$ is an increasing sequence of functions converging pointwise to f then $\mu([f > t]) = \lim_{n\to\infty} \mu([f_n > t])$.*

Proof: The sets, $[f_n > t]$ are increasing and their union is $[f > t]$ because if $f(\omega) > t$, then for all n large enough, $f_n(\omega) > t$ also. Therefore, the desired conclusion follows from properties of measures, the one which says that if $E_n \uparrow E$, then $\mu(E_n) \uparrow \mu(E)$. $\blacksquare$

Note how it was important to have strict inequality in the definition.

Lemma 10.11.3 *Suppose $s \geq 0$ is a simple function, $s(\omega) \equiv \sum_{k=1}^{n} a_k \mathcal{X}_{E_k}(\omega)$ where the a_k are the distinct nonzero values of $s, 0 < a_1 < a_2 < \cdots < a_n$ on the measurable sets E_k. Suppose ϕ is a C^1 function defined on $[0,\infty)$ which has the property that $\phi(0) = 0, \phi'(t) > 0$ for all t. Then*

$$\int_0^\infty \phi'(t)\,\mu([s > t])\,dm(t) = \int \phi(s)\,d\mu(s).$$

Proof: First note that if $\mu(E_k) = \infty$ for any k then both sides equal ∞ and so without loss of generality, assume $\mu(E_k) < \infty$ for all k. Letting $a_0 \equiv 0$, the left side equals

$$\sum_{k=1}^{n} \int_{a_{k-1}}^{a_k} \phi'(t)\,\mu([s > t])\,dm(t) = \sum_{k=1}^{n} \int_{a_{k-1}}^{a_k} \phi'(t) \sum_{i=k}^{n} \mu(E_i)\,dm$$

$$= \sum_{k=1}^{n} \sum_{i=k}^{n} \mu(E_i) \int_{a_{k-1}}^{a_k} \phi'(t)\,dm = \sum_{k=1}^{n} \sum_{i=k}^{n} \mu(E_i)\,(\phi(a_k) - \phi(a_{k-1}))$$

$$= \sum_{i=1}^{n} \mu(E_i) \sum_{k=1}^{i} (\phi(a_k) - \phi(a_{k-1})) = \sum_{i=1}^{n} \mu(E_i)\,\phi(a_i) = \int \phi(s)\,d\mu. \blacksquare$$

With this lemma the next theorem which is the main result follows easily.

Theorem 10.11.4 *Let $f \geq 0$ be measurable and let ϕ be a C^1 function defined on $[0,\infty)$ which satisfies $\phi'(t) > 0$ for all $t > 0$ and $\phi(0) = 0$. Then*

$$\int \phi(f)\,d\mu = \int_0^\infty \phi'(t)\,\mu([f > t])\,dm.$$

Proof: By Theorem 9.1.6 on Page 244 there exists an increasing sequence of nonnegative simple functions, $\{s_n\}$ which converges pointwise to f. By the monotone convergence theorem and Lemma 10.11.2,

$$\int \phi(f)\,d\mu = \lim_{n\to\infty} \int \phi(s_n)\,d\mu = \lim_{n\to\infty} \int_0^\infty \phi'(t)\,\mu([s_n > t])\,dm$$

$$= \int_0^\infty \phi'(t)\,\mu([f > t])\,dm \blacksquare$$

This theorem can be generalized to a situation in which ϕ is only increasing and continuous. In the generalization I will replace the symbol ϕ with F to coincide with earlier notation.

The following lemma and theorem say essentially that for F an increasing function equal to 0 at 0,

$$\int_{(0,\infty]} \mu([f > t])\,dF = \int_\Omega F(f)\,d\mu$$

I think it is particularly memorable if F is differentiable when it looks like what was just discussed.

$$\int_{(0,\infty]} \mu([f > t])\,F'(t)\,dt = \int_\Omega F(f)\,d\mu$$

Lemma 10.11.5 *Suppose $s \geq 0$ is a measurable simple function,*

$$s(\omega) \equiv \sum_{k=1}^{n} a_k \mathcal{X}_{E_k}(\omega)$$

where the a_k are the distinct nonzero values of $s, a_1 < a_2 < \cdots < a_n$. Suppose F is an increasing function defined on $[0,\infty), F(0) = 0, F$ being continuous at 0 from the right and continuous at every a_k. Then letting μ be a measure and $(\Omega, \mathscr{F}, \mu)$ a measure space,

$$\int_{(0,\infty]} \mu([s > t])\, d\nu = \int_\Omega F(s)\, d\mu.$$

where the integral on the left is the Lebesgue integral for the Lebesgue Stieltjes measure ν which comes from the increasing function F as in Theorem 10.10.1 above.

Proof: This follows from the following computation. Since F is continuous at 0 and the values a_k,

$$\int_0^\infty \mu([s > t])\, d\nu(t) = \sum_{k=1}^n \int_{(a_{k-1},a_k]} \mu([s > t])\, d\nu(t)$$

$$= \sum_{k=1}^n \int_{(a_{k-1},a_k]} \sum_{j=k}^n \mu(E_j)\, dF(t) = \sum_{j=1}^n \mu(E_j) \sum_{k=1}^j \nu((a_{k-1}, a_k])$$

$$= \sum_{j=1}^n \mu(E_j) \sum_{k=1}^j (F(a_k) - F(a_{k-1})) = \sum_{j=1}^n \mu(E_j) F(a_j) \equiv \int_\Omega F(s)\, d\mu \blacksquare$$

Now here is the generalization to nonnegative measurable f.

Theorem 10.11.6 *Let $f \geq 0$ be measurable with respect to $\mathscr{F}$, $(\Omega, \mathscr{F}, \mu)$ a measure space, and let F be an increasing continuous function defined on $[0,\infty)$ and $F(0) = 0$. Then*

$$\int_\Omega F(f)\, d\mu = \int_{(0,\infty]} \mu([f > t])\, d\nu(t)$$

where ν is the Lebesgue Stieltjes measure determined by F as in Theorem 10.10.1 above.

Proof: By Theorem 9.1.6 on Page 244 there exists an increasing sequence of nonnegative simple functions, $\{s_n\}$ which converges pointwise to f. By the monotone convergence theorem and Lemma 10.11.5,

$$\begin{aligned} \int_\Omega F(f)\, d\mu &= \lim_{n\to\infty} \int_\Omega F(s_n)\, d\mu = \lim_{n\to\infty} \int_{(0,\infty]} \mu([s_n > t])\, d\nu \\ &= \int_{(0,\infty]} \mu([f > t])\, d\nu \blacksquare \end{aligned}$$

Note that the function $t \to \mu([f > t])$ is a decreasing function. Therefore, one can make sense of an improper Riemann Stieltjes integral $\int_0^\infty \mu([f > t])\, dF(t)$. With more work, one can have this equal to the corresponding Lebesgue integral above.

10.12 Good Lambda Inequality

There is a very interesting and important inequality called the good lambda inequality (I am not sure if there is a bad lambda inequality.) which follows from the above theory of distribution functions. It involves the inequality

$$\mu([f > \beta\lambda] \cap [g \leq \delta\lambda]) \leq \phi(\delta)\, \mu([f > \lambda]) \tag{10.13}$$

for $\beta > 1$, nonnegative functions f, g and is supposed to hold for all small positive δ and $\phi(\delta) \to 0$ as $\delta \to 0$. Note the left side is small when g is large and f is small. The inequality involves dominating an integral involving f with one involving g as described below. As above, ν is the Lebesgue Stieltjes measure described above in terms of F, an increasing function. Is there any way to see the inequality in 10.13 might make sense? Look at the expression on the left. If δ is small enough, you might think that the intersection of the two sets would have smaller measure than $\mu([f > \lambda])$. The inequality says that this is the case in a precise manner.

Theorem 10.12.1 *Let $(\Omega, \mathcal{F}, \mu)$ be a finite measure space and let F be a continuous increasing function defined on $[0,\infty)$ such that $F(0) = 0$. Suppose also that for every $\alpha > 1$, there exists a constant C_α such that for all $x \in [0,\infty)$, $F(\alpha x) \leq C_\alpha F(x)$. Also suppose f, g are nonnegative measurable functions and there exists $\beta > 1, 0 < r \leq 1$, such that for all $\lambda > 0$ and $1 > \delta > 0$,*

$$\mu([f > \beta\lambda] \cap [g \leq r\delta\lambda]) \leq \phi(\delta)\mu([f > \lambda]) \tag{10.14}$$

where $\lim_{\delta \to 0+} \phi(\delta) = 0$ and ϕ is increasing. Under these conditions, there exists a constant C depending only on β, ϕ, r such that

$$\int_\Omega F(f(\omega))\, d\mu(\omega) \leq C \int_\Omega F(g(\omega))\, d\mu(\omega).$$

Proof: Let $\beta > 1$ be as given above. First suppose f is bounded. This is so there can be no question of existence of the integrals. $\int_\Omega F(f)\, d\mu = \int_\Omega F\left(\beta \frac{f}{\beta}\right) d\mu \leq C_\beta \int_\Omega F\left(\frac{f}{\beta}\right) d\mu$. Recalling that ν is the Lebesgue Stieltjes measure which comes from F, it follows from Theorem 10.11.6,

$$= C_\beta \int_0^\infty \mu([f/\beta > \lambda])\, d\nu = C_\beta \int_0^\infty \mu([f > \beta\lambda])\, d\nu$$

Now using the given inequality,

$$\begin{aligned} &= C_\beta \int_0^\infty \mu([f > \beta\lambda] \cap [g \leq r\delta\lambda])\, d\nu + C_\beta \int_0^\infty \mu([f > \beta\lambda] \cap [g > r\delta\lambda])\, d\nu \\ &\leq C_\beta \phi(\delta) \int_0^\infty \mu([f > \lambda])\, d\nu + C_\beta \int_0^\infty \mu([g > r\delta\lambda])\, d\nu \\ &\leq C_\beta \phi(\delta) \int_\Omega F(f)\, d\mu + C_\beta \int_\Omega F\left(\frac{g}{r\delta}\right) d\mu \end{aligned}$$

Now choose δ small enough that $C_\beta \phi(\delta) < \frac{1}{2}$ and then subtract the first term on the right in the above from both sides. It follows from the properties of F again that

$$\frac{1}{2} \int_\Omega F(f)\, d\mu \leq C_\beta C_{(r\delta)^{-1}} \int_\Omega F(g)\, d\mu.$$

This establishes the inequality in the case where f is bounded.

In general, let $f_n = \min(f, n)$. For $n \leq \lambda$, the inequality

$$\mu([f > \beta\lambda] \cap [g \leq r\delta\lambda]) \leq \phi(\delta)\mu([f > \lambda])$$

holds with f replaced with f_n because both sides equal 0 thanks to $\beta > 1$. If $n > \lambda$, then $[f > \lambda] = [f_n > \lambda]$ and so the inequality still holds because in this case,

$$\begin{aligned} \mu([f_n > \beta\lambda] \cap [g \leq r\delta\lambda]) &\leq \mu([f > \beta\lambda] \cap [g \leq r\delta\lambda]) \\ &\leq \phi(\delta)\mu([f > \lambda]) = \phi(\delta)\mu([f_n > \lambda]) \end{aligned}$$

Therefore, 10.14 is valid with f replaced with f_n. Now pass to the limit as $n \to \infty$ and use the monotone convergence theorem. ■

10.13 Radon Nikodym Theorem

Let μ, ν be two finite measures on the measurable space $(\Omega, \mathscr{F})$ and let $\alpha \geq 0$. Let $\lambda \equiv \nu - \alpha\mu$. Then it is clear that if $\{E_i\}_{i=1}^{\infty}$ are disjoint sets of $\mathscr{F}$, then $\lambda(\cup_i E_i) = \sum_{i=1}^{\infty} \lambda(E_i)$ and that the series converges. The next proposition is fairly obvious.

Proposition 10.13.1 *Let $(\Omega, \mathscr{F}, \lambda)$ be a measure space and let $\lambda : \mathscr{F} \to [0, \infty)$ be a measure. Then λ is a finite measure.*

Proof: Since $\lambda(\Omega) < \infty$ this is a finite measure. ■

Definition 10.13.2 *Let $(\Omega, \mathscr{F})$ be a measurable space and let $\lambda : \mathscr{F} \to \mathbb{R}$ satisfy: If $\{E_i\}_{i=1}^{\infty}$ are disjoint sets of $\mathscr{F}$, then $\lambda(\cup_i E_i) = \sum_{i=1}^{\infty} \lambda(E_i)$ and the series converges. Such a real valued function is called a signed measure. In this context, a set $E \in \mathscr{F}$ is called positive if whenever F is a measurable subset of E, it follows $\lambda(F) \geq 0$. A negative set is defined similarly. Note that this requires $\lambda(\Omega) \in \mathbb{R}$.*

Lemma 10.13.3 *The countable union of disjoint positive sets is positive.*

Proof: Let E_i be positive and consider $E \equiv \cup_{i=1}^{\infty} E_i$. If $A \subseteq E$ with A measurable, then $A \cap E_i \subseteq E_i$ and so $\lambda(A \cap E_i) \geq 0$. Hence $\lambda(A) = \sum_i \lambda(A \cap E_i) \geq 0$. ■

Lemma 10.13.4 *Let λ be a signed measure on $(\Omega, \mathscr{F})$. If $E \in \mathscr{F}$ with $0 < \lambda(E)$, then E has a measurable subset which is positive.*

Proof: If every measurable subset F of E has $\lambda(F) \geq 0$, then E is positive and we are done. Otherwise there exists measurable $F \subseteq E$ with $\lambda(F) < 0$. Let the elements of $\mathfrak{F}$ consist of sets of disjoint sets of measurable subsets of E each of which has measure less than 0. Partially order $\mathfrak{F}$ by set inclusion. By the Hausdorff maximal theorem, there is a maximal chain $\mathscr{C}$. Then $\cup\mathscr{C}$ is a set consisting of disjoint measurable sets $F \in \mathscr{F}$ such that $\lambda(F) < 0$. Since each set in $\cup\mathscr{C}$ has measure strictly less than 0, it follows that $\cup\mathscr{C}$ is a countable set, $\{F_i\}_{i=1}^{\infty}$. Otherwise, there would exist an infinite subset of $\cup\mathscr{C}$ with each set having measure less than $-\frac{1}{n}$ for some $n \in \mathbb{N}$ so λ would not be real valued. Letting $F = \cup_i F_i$, then $E \setminus F$ has no measurable subsets S for which $\lambda(S) < 0$ since, if it did, $\mathscr{C}$ would not have been maximal. Thus $E \setminus F$ is positive. ■

A major result is the following, called a Hahn decomposition.

Theorem 10.13.5 *Let λ be a signed measure on a measurable space $(\Omega, \mathscr{F})$. Then there are disjoint measurable sets P, N such that P is a positive set, N is a negative set, and $P \cup N = \Omega$.*

Proof: If Ω is either positive or negative, there is nothing to show, so suppose Ω is neither positive nor negative. $\mathfrak{F}$ will consist of collections of disjoint measurable sets F such that $\lambda(F) > 0$. Thus each element of $\mathfrak{F}$ is necessarily countable. Partially order $\mathfrak{F}$ by set inclusion and use the Hausdorff maximal theorem to get $\mathscr{C}$ a maximal chain. Then, as in the above lemma, $\cup\mathscr{C}$ is countable, say $\{P_i\}_{i=1}^{\infty}$ because $\lambda(F) > 0$ for each $F \in \cup\mathscr{C}$ and λ has values in $\mathbb{R}$. The sets in $\cup\mathscr{C}$ are disjoint because if A, B are two of them, then they are both in a single element of $\mathscr{C}$. Letting $P \equiv \cup_i P_i$, and $N = P^C$, it follows from Lemma 10.13.3 that P is positive. It is also the case that N must be negative because otherwise, $\mathscr{C}$ would not be maximal. ■

Clearly a Hahn decomposition is not unique. For example, you could have obtained a different Hahn decomposition if you had considered disjoint negative sets F for which $\lambda(F) < 0$ in the above argument .

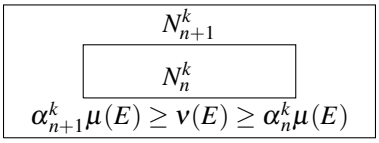
Let $k \in \mathbb{N}$, $\{\alpha_n^k\}_{n=0}^{\infty}$ be equally spaced points $\alpha_n^k = 2^{-k}n$. Then $\alpha_{2n}^k = 2^{-k}(2n) = 2^{-(k-1)}n \equiv \alpha_n^{k-1}$ and $\alpha_{2n}^{k+1} \equiv 2^{-(k+1)}2n = \alpha_n^k$. Similarly $N_{2n}^{k+1} = N_n^k$ because these depend on the α_n^k. Also let $\left(P_n^k, N_n^k\right)$ be a Hahn decomposition for the signed measure $\nu - \alpha_n^k\mu$ where ν, μ are two finite measures. Now from the definition, $N_{n+1}^k \setminus N_n^k = N_{n+1}^k \cap P_n^k$. Also, $N_n \subseteq N_{n+1}$ for each n and we can take $N_0 = \emptyset$. then $\left\{N_{n+1}^k \setminus N_n^k\right\}_{n=0}^{\infty}$ covers all of Ω except possibly for a set of μ measure 0.

Lemma 10.13.6 *Let $S \equiv \Omega \setminus \left(\cup_n N_n^k\right) = \Omega \setminus \left(\cup_n N_n^l\right)$ for any l. Then $\mu(S) = 0$.*

Proof: $S = \cap_n P_n^k$ so for all n, $\nu(S) - \alpha_n^k\mu(S) \geq 0$. But letting $n \to \infty$, it must be that $\mu(S) = 0$. ■

As just noted, if $E \subseteq N_{n+1}^k \setminus N_n^k$, then

$$\nu(E) - \alpha_n^k\mu(E) \geq 0 \geq \nu(E) - \alpha_{n+1}^k\mu(E), \text{ so } \alpha_{n+1}^k\mu(E) \geq \nu(E) \geq \alpha_n^k\mu(E) \quad (10.15)$$



Then define $f^k(\omega) \equiv \sum_{n=0}^{\infty} \alpha_n^k \mathscr{X}_{\Delta_n^k}(\omega)$ where $\Delta_m^k \equiv N_{m+1}^k \setminus N_m^k$. Thus,

$$f^k = \sum_{n=0}^{\infty} \alpha_{2n}^{k+1} \mathscr{X}_{\left(N_{2n+2}^{k+1} \setminus N_{2n}^{k+1}\right)} = \sum_{n=0}^{\infty} \alpha_{2n}^{k+1} \mathscr{X}_{\Delta_{2n+1}^{k+1}} + \sum_{n=0}^{\infty} \alpha_{2n}^{k+1} \mathscr{X}_{\Delta_{2n}^{k+1}}$$

$$\leq \sum_{n=0}^{\infty} \alpha_{2n+1}^{k+1} \mathscr{X}_{\Delta_{2n+1}^{k+1}} + \sum_{n=0}^{\infty} \alpha_{2n}^{k+1} \mathscr{X}_{\Delta_{2n}^{k+1}} = f^{k+1} \quad (10.16)$$

Thus $k \to f^k(\omega)$ is increasing. Let $f(\omega) \equiv \lim_{k\to\infty} f^k(\omega)$. Thus $f = 0$ on S. Now let E be measurable. Thus $\mu(E) = \mu(E \cap S) + \mu\left(E \cap S^C\right)$, similar for λ and $\lambda\left(E \cap S^C\right) = \sum_n \lambda\left(E \cap S^C \cap \Delta_n^k\right)$. To save space, let $\tilde{E} \equiv E \cap S^C$. Then using 10.15

$$\int \mathscr{X}_{\tilde{E}} f^k d\mu \leq \sum_{n=0}^{\infty} \alpha_{n+1}^k \mu\left(\tilde{E} \cap \Delta_n^k\right) \leq \sum_{n=0}^{\infty} \alpha_n^k \mu\left(\tilde{E} \cap \Delta_n^k\right) + \sum_{n=0}^{\infty} 2^{-k} \mu\left(\tilde{E} \cap \Delta_n^k\right)$$

$$\leq \sum_{n=0}^{\infty} \nu\left(\tilde{E} \cap \Delta_n^k\right) + 2^{-k}\mu\left(\tilde{E}\right) = \nu\left(\tilde{E}\right) + 2^{-k}\mu\left(\tilde{E}\right) \leq \int \mathscr{X}_{\tilde{E}} f^k d\mu + 2^{-k}\mu\left(\tilde{E}\right) \quad (10.17)$$

From the monotone convergence theorem it follows $\nu\left(\tilde{E}\right)=\int \mathcal{X}_{\tilde{E}} f d\mu=\int \mathcal{X}_E f d\mu$.

This proves most of the following theorem which is the Radon Nikodym theorem.

Theorem 10.13.7 *Let ν and μ be finite measures defined on a measurable space $(\Omega,\mathcal{F})$. Then there exists a measurable set S with $\mu(S)=0$ and a nonnegative measurable function $\omega\to f(\omega)$ such that $\nu(E)=\int_E f d\mu+\nu(E\cap S)$.*

Proof: Let S be defined in Lemma 10.13.6 so $S\equiv\Omega\setminus\left(\cup_n N_n^k\right)$ and $\mu(S)=0$. If $E\in\mathcal{F}$, and f as described above,

$$\nu(E)=\nu\left(E\cap S^C\right)+\nu(E\cap S)=\int_{E\cap S^C} f d\mu+\nu(E\cap S)=\int_E f d\mu+\nu(E\cap S)$$

Thus if $E\subseteq S^C$, we have $\nu(E)=\int_E f d\mu$. ■

Definition 10.13.8 *Let μ,ν be finite measures on $(\Omega,\mathcal{F})$. Then $\nu\ll\mu$ means that whenever $\mu(E)=0$, it follows that $\nu(E)=0$.*

Sometimes people write $f=\frac{d\lambda}{d\mu}$, in the case $\nu\ll\mu$ and this is called the Radon Nikodym derivative.

Proposition 10.13.9 *If ν,μ are finite measures and $\nu\ll\mu$, then there exists nonnegative measurable f such that $\nu(E)=\int_E f d\mu$.*

Proof: In Theorem 10.13.7, $\nu(E\cap S)=0$ because $\mu(S)=0$ and $\nu\ll\mu$. ■

Definition 10.13.10 *Let S be in the above theorem. Then*

$$\nu_{||}(E)\equiv\nu\left(E\cap S^C\right)=\int_{E\cap S^C} f d\mu=\int_E f d\mu$$

while $\nu_\perp(E)\equiv\nu(E\cap S)$. Thus $\nu_{||}\ll\mu$ and $\nu_\perp$ is nonzero only on sets which are contained in S which has μ measure 0.

Corollary 10.13.11 *In the above situation, let λ be a signed measure and let $\lambda\ll\mu$ meaning that if $\mu(E)=0\Rightarrow\lambda(E)=0$. Here assume that μ is a finite measure. Then there exists $h\in L^1$ such that $\lambda(E)=\int_E h d\mu$.*

Proof: Let $P\cup N$ be a Hahn decomposition of λ. Let

$$\lambda_+(E)\equiv\lambda(E\cap P),\ \ \lambda_-(E)\equiv-\lambda(E\cap N).$$

Then both λ_+ and λ_- are absolutely continuous measures and so there are nonnegative h_+ and h_- with $\lambda_-(E)=\int_E h_- d\mu$ and a similar equation for λ_+. Then $0\le-\lambda(\Omega\cap N)\le\lambda_-(\Omega)<\infty$, similar for λ_+ so both of these measures are necessarily finite. Hence both h_- and h_+ are in L^1 so $h\equiv h_+-h_-$ is also in L^1 and $\lambda(E)=\lambda_+(E)-\lambda_-(E)=\int_E(h_+-h_-)d\mu$. ■

Proposition 10.13.12 *This Lebesgue decomposition is unique. If $f,\hat{f}$ both work in Theorem 10.13.7, then $f=\hat{f}$ μ a.e. This function $f\in L^1(\Omega)$, $\int_\Omega f d\mu<\infty$.*

Proof: Say $\nu_{||}+\nu_{\perp}=\hat{\nu}_{||}+\hat{\nu}_{\perp}$. Then $\nu_{||}(E)-\hat{\nu}_{||}(E)=\nu_{\perp}(E)-\hat{\nu}_{\perp}(E)$. If $\mu(E)=0$, then the left side is also 0 and so $\nu_{\perp}(E)-\hat{\nu}_{\perp}(E)=0$. But then for any E,

$$\nu_{\perp}(E)-\hat{\nu}_{\perp}(E)=\int_E h d\mu \tag{10.18}$$

for h a function in $L^1(\Omega)$. This is because $\nu_{\perp}-\hat{\nu}_{\perp}$ is a signed measure $\lambda \ll \mu$ and Corollary 10.13.11. From the above, if $S,\hat{S}$ are the exceptional sets of μ measure zero,

$$\nu_{\perp}(E)-\hat{\nu}_{\perp}(E)=\nu_{\perp}\left(E\cup\left(S\cup\hat{S}\right)\right)-\hat{\nu}_{\perp}\left(E\cup\left(S\cup\hat{S}\right)\right)=\int_{E\cup\left(S\cup\hat{S}\right)} h d\mu=0 \tag{10.19}$$

because $\mu\left(S\cup\hat{S}\right)=0$ and so the right side of 10.18 must be 0 after all, so $\nu_{\perp}(E)=\hat{\nu}_{\perp}(E)$. It follows that $\nu_{||}=\hat{\nu}_{||}$ also. Now in Theorem 10.13.7, if you have two $f,\hat{f}$ which work, then

$$\nu_{||}(E)=\int_E f d\mu=\int_E \hat{f} d\mu=\hat{\nu}_{||}(E) \tag{10.20}$$

and so, $f=\hat{f}$ a.e. because you can apply this equation to $E_n\equiv\left[f-\hat{f}>1/n\right]$ and conclude that

$$0=\int_{E_n} f-\hat{f} d\mu\geq\frac{1}{n}\mu(E_n)=0 \tag{10.21}$$

so $\mu\left(\left[f-\hat{f}>0\right]\right)=\cup_m\mu(E_n)=0$. Similarly $\mu\left(\left[\hat{f}-f>0\right]\right)=0$. ■

This unique decomposition of a measure ν into the sum of two measures, one absolutely continuous with respect to μ and the other supported on a set of μ measure zero is called the Lebesgue decomposition.

Definition 10.13.13 *A measure space $(\Omega,\mathcal{F},\mu)$ is σ finite if there are countably many measurable sets $\{\Omega_n\}$ such that μ is finite on measurable subsets of Ω_n.*

There is a routine corollary of the above theorem.

Corollary 10.13.14 *Suppose μ,ν are both σ finite measures defined on $(\Omega,\mathcal{F})$. Then a similar conclusion to the above theorem can be obtained.*

$$\nu(E)=\int_E f d\mu+\nu(E\cap S),\ \mu(S)=0 \tag{10.22}$$

for f a nonnegative measurable function. If $\nu(\Omega)<\infty$, then $f\in L^1(\Omega)$. This f is unique up to a set of μ measure zero.

Proof: Since both μ,ν are σ finite, there are $\left\{\tilde{\Omega}_k\right\}_{k=1}^{\infty}$ such that $\nu\left(\tilde{\Omega}_k\right),\mu\left(\tilde{\Omega}_k\right)$ are finite. Let $\Omega_0=\emptyset$ and $\Omega_k\equiv\tilde{\Omega}_k\setminus\left(\cup_{j=0}^{k-1}\tilde{\Omega}_j\right)$ so that μ,ν are finite on Ω_k and the Ω_k are disjoint. Let $\mathcal{F}_k$ be the measurable subsets of Ω_k, equivalently the intersections with Ω_k with sets of $\mathcal{F}$. Now let $\nu_k(E)\equiv\nu(E\cap\Omega_k)$, similar for μ_k. By Theorem 10.13.7, there exists $S_k\subseteq\Omega_k$, and f_k as described there, unique up to sets of μ measure 0. Thus $\mu_k(S_k)=0$ and

$$\nu_k(E)=\int_{E\cap\Omega_k} f_k d\mu_k+\nu_k(E\cap S_k)$$

Now let $f(\omega) \equiv f_k(\omega)$ for $\omega \in \Omega_k$. Thus

$$\nu(E \cap \Omega_k) = \nu(E \cap (\Omega_k \setminus S_k)) + \nu(E \cap S_k) = \int_{E \cap \Omega_k} f d\mu + \nu(E \cap S_k) \tag{10.23}$$

Summing over all k,

$$\nu(E) = \nu\left(E \cap S^C\right) + \nu(E \cap S) = \int_E f d\mu + \nu(E \cap S).$$

In particular, if $\nu \ll \mu$, then $\nu(E \cap S) = 0$ and so $\nu(E) = \int_E f d\mu$. The last claim is obvious from 10.22. ■

10.14 Abstract Product Measures and Integrals

You can take the product measure for any two finite or σ finite measures. This is described here.

Definition 10.14.1 *A measure space $(\Omega, \mathcal{F}, \mu)$ is called σ finite if there are measurable subsets Ω_n such that $\mu(\Omega_n) < \infty$ and $\Omega = \cup_{n=1}^{\infty} \Omega_n$. σ finite*

Given two finite measure spaces, $(X, \mathcal{F}, \mu)$ and $(Y, \mathcal{S}, \nu)$, there is a way to define a σ algebra of subsets of $X \times Y$, denoted by $\mathcal{F} \times \mathcal{S}$ and a measure, denoted by $\mu \times \nu$ defined on this σ algebra such that

$$(\mu \times \nu)(A \times B) = \mu(A)\nu(B)$$

whenever $A \in \mathcal{F}$ and $B \in \mathcal{S}$. This σ algebra $\mathcal{F} \times \mathcal{S}$ is the smallest σ algebra which contains all sets of the form $A \times B$ where $A \in \mathcal{F}$ and $B \in \mathcal{S}$. These sets in $\mathcal{F} \times \mathcal{S}$ are called the product measurable sets.

Definition 10.14.2 *Let $(X, \mathcal{F}, \mu)$ and $(Y, \mathcal{S}, \nu)$ be two measure spaces. A measurable rectangle is a set of the form $A \times B$ where $A \in \mathcal{F}$ and $B \in \mathcal{S}$.*

We also have the following important proposition which holds in every context independent of any measure.

Proposition 10.14.3 *Let $E \subseteq \mathcal{F} \times \mathcal{S}$ be product measurable $\mathcal{F} \times \mathcal{S}$ where $\mathcal{F}$ is a σ algebra of sets of X and $\mathcal{S}$ is a σ algebra of sets of $\mathcal{S}$. then if $E_x \equiv \{y \in Y : (x,y) \in E\}$ and $E_y \equiv \{x \in X : (x,y) \in E\}$, then $E_x \in \mathcal{F}$ and $E_y \in \mathcal{S}$.*

Proof: It is obvious that if $\mathcal{K}$ is the measurable rectangles, then the conclusion of the proposition holds. If $\mathcal{G}$ consists of the sets of $\mathcal{F} \times \mathcal{S}$ for which the proposition holds, then it is clearly closed with respect to countable disjoint unions and complements. This is obvious in the case of a countable disjoint union since $\left(\cup_i E^i\right)_x = \cup_i E^i_x$, similar for y. As to complement, if $E \in \mathcal{G}$, then $E_x \in \mathcal{S}$ and so $\left(E^C\right)_x = (E_x)^C \in \mathcal{S}$. It is similar for y. By Dynkin's lemma, $\mathcal{G} \supseteq \mathcal{F} \times \mathcal{S}$. However $\mathcal{G}$ was defined as a subset of $\mathcal{F} \times \mathcal{S}$ so these are equal. ■

Let $(X, \mathcal{F}, \mu)$ and $(Y, \mathcal{S}, \nu)$ be two finite measure spaces. Define $\mathcal{K}$ to be the set of measurable rectangles, $A \times B$, $A \in \mathcal{F}$ and $B \in \mathcal{S}$. Let

$$\mathcal{G} \equiv \left\{ E \subseteq X \times Y : \int_Y \int_X \mathcal{X}_E d\mu d\nu = \int_X \int_Y \mathcal{X}_E d\nu d\mu \right\} \tag{10.24}$$

where in the above, part of the requirement is for all integrals to make sense.

Then $\mathscr{K} \subseteq \mathscr{G}$. This is obvious.

Next I want to show that if $E \in \mathscr{G}$ then $E^C \in \mathscr{G}$. Observe $\mathscr{X}_{E^C} = 1 - \mathscr{X}_E$ and so

$$\begin{aligned} \int_Y \int_X \mathscr{X}_{E^C} d\mu d\nu &= \int_Y \int_X (1 - \mathscr{X}_E)\, d\mu d\nu = \int_X \int_Y (1 - \mathscr{X}_E)\, d\nu d\mu \\ &= \int_X \int_Y \mathscr{X}_{E^C} d\nu d\mu \end{aligned}$$

which shows that if $E \in \mathscr{G}$, then $E^C \in \mathscr{G}$.

Next I want to show $\mathscr{G}$ is closed under countable unions of disjoint sets of $\mathscr{G}$. Let $\{A_i\}$ be a sequence of disjoint sets from $\mathscr{G}$. Then, using the monotone convergence theorem as needed,

$$\begin{aligned} \int_Y \int_X \mathscr{X}_{\cup_{i=1}^\infty A_i} d\mu d\nu &= \int_Y \int_X \sum_{i=1}^\infty \mathscr{X}_{A_i} d\mu d\nu = \int_Y \sum_{i=1}^\infty \int_X \mathscr{X}_{A_i} d\mu d\nu \\ &= \sum_{i=1}^\infty \int_Y \int_X \mathscr{X}_{A_i} d\mu d\nu = \sum_{i=1}^\infty \int_X \int_Y \mathscr{X}_{A_i} d\nu d\mu \\ &= \int_X \sum_{i=1}^\infty \int_Y \mathscr{X}_{A_i} d\nu d\mu = \int_X \int_Y \sum_{i=1}^\infty \mathscr{X}_{A_i} d\nu d\mu \\ &= \int_X \int_Y \mathscr{X}_{\cup_{i=1}^\infty A_i} d\nu d\mu, \end{aligned} \tag{10.25}$$

Thus $\mathscr{G}$ is closed with respect to countable disjoint unions.

From Lemma 9.3.2, $\mathscr{G} \supseteq \sigma(\mathscr{K})$, the smallest σ algebra containing $\mathscr{K}$. Also the computation in 10.25 implies that on $\sigma(\mathscr{K})$ one can define a measure, denoted by $\mu \times \nu$ and that for every $E \in \sigma(\mathscr{K})$,

$$(\mu \times \nu)(E) = \int_Y \int_X \mathscr{X}_E d\mu d\nu = \int_X \int_Y \mathscr{X}_E d\nu d\mu. \tag{10.26}$$

with each iterated integral making sense.

Next is product measure. First is the case of finite measures. Then this will extend to σ finite measures. The following theorem is Fubini's theorem.

Theorem 10.14.4 *Let $f : X \times Y \to [0,\infty]$ be measurable with respect to the σ algebra, $\sigma(\mathscr{K}) \equiv \mathscr{F} \times \mathscr{S}$ just defined and let $\mu \times \nu$ be the product measure of 10.26 where μ and ν are finite measures on $(X,\mathscr{F})$ and $(Y,\mathscr{S})$ respectively. Then*

$$\int_{X\times Y} f d(\mu \times \nu) = \int_Y \int_X f d\mu d\nu = \int_X \int_Y f d\nu d\mu.$$

Proof: Let $\{s_n\}$ be an increasing sequence of $\sigma(\mathscr{K}) \equiv \mathscr{F} \times \mathscr{S}$ measurable simple functions which converges pointwise to f. The above equation holds for s_n in place of f from what was shown above. The final result follows from passing to the limit and using the monotone convergence theorem. ■

Of course one can generalize right away to measures which are only σ finite. This is also called Fubini's theorem.

Definition 10.14.5 *Let $(X,\mathscr{F},\mu),(Y,\mathscr{S},\nu)$ both be σ finite. Thus there exist disjoint measurable X_n with $\cup_{n=1}^{\infty}X_n$ and disjoint measurable Y_n with $\cup_{n=1}^{\infty}Y_n=Y$ such that μ,ν restricted to X_n,Y_n respectively are finite measures. Let $\mathscr{F}_n$ be intersections of sets of $\mathscr{F}$ with X_n and $\mathscr{S}_n$ similarly defined. Then letting $\mathscr{K}$ consist of all measurable rectangles $A\times B$ for $A\in\mathscr{F},B\in\mathscr{S}$, and letting $\mathscr{F}\times\mathscr{S}\equiv\sigma(\mathscr{K})$ define the product measure of E contained in this σ algebra as $(\mu\times\nu)(E)\equiv\sum_n\sum_m(\mu_n\times\nu_m)(E\cap(X_n\times Y_m))$.*

Lemma 10.14.6 *The above definition yields a well defined measure on $\mathscr{F}\times\mathscr{S}$.*

Proof: This follows from the standard theorems about sums of nonnegative numbers. See Theorem 2.5.4. For example if you have two other disjoint sequences X_k,Y_l on which the measures are finite, then

$$\begin{aligned}(\mu\times\nu)(E) &= \sum_n\sum_m\sum_k\sum_l(\mu_n\times\nu_m)(E\cap(X_n\cap X_k\times Y_m\cap Y_l))\\ &= \sum_k\sum_l\sum_n\sum_m(\mu_k\times\nu_l)(E\cap(X_n\cap X_k\times Y_m\cap Y_l))\end{aligned}$$

and so the definition with respect to the two different increasing sequences gives the same thing. Thus the definition is well defined. $(\mu\times\nu)$ is a measure because if the E_i are disjoint $\mathscr{F}\times\mathscr{S}$ measurable sets and $E=\cup_iE_i$,

$$(\mu\times\nu)(E)\equiv\sum_n\sum_m(\mu_n\times\nu_m)(\cup_iE_i\cap(X_n\times Y_m))=\sum_n\sum_m\sum_i(\mu_n\times\nu_m)(E_i\cap(X_n\times Y_m))$$

$$=\sum_i\sum_n\sum_m(\mu_n\times\nu_m)(E_i\cap(X_n\times Y_m))\equiv\sum_i(\mu\times\nu)(E_i)\ \blacksquare$$

Theorem 10.14.7 *Let $f:X\times Y\to[0,\infty]$ be measurable with respect to the σ algebra, $\sigma(\mathscr{K})$ just defined as the smallest σ algebra containing the measurable rectangles, and let $\mu\times\nu$ be the product measure of 10.26 where μ and ν are σ finite measures on $(X,\mathscr{F})$ and $(Y,\mathscr{S})$ respectively. (10.14.1) Then*

$$\int_{X\times Y}fd(\mu\times\nu)=\int_Y\int_Xfd\mu d\nu=\int_X\int_Yfd\nu d\mu. \tag{10.27}$$

Proof: Letting $E\in\mathscr{F}\times\mathscr{S}$,

$$\int_{X\times Y}\mathscr{X}_Ed(\mu\times\nu)\equiv(\mu\times\nu)(E)\equiv\sum_n\sum_m(\mu_n\times\nu_m)(E\cap(X_n\times Y_m))$$

$$=\sum_n\sum_m\int_{Y_n}\int_{X_n}\mathscr{X}_Ed\mu_nd\nu_n=\int_Y\int_X\mathscr{X}_Ed\mu d\nu$$

the last coming from a use of the monotone convergence theorem applied to sums. It follows that 10.27 holds for simple functions and then from monotone convergence theorem and Theorem 9.1.6, it holds for nonnegative $\mathscr{F}\times\mathscr{S}$ measurable functions. $\blacksquare$

It is also useful to note that all the above holds for $\prod_{i=1}^pX_i$ in place of $X\times Y$ and μ_i a measure on $\mathscr{F}_i$ a σ algebra of sets of X_i. You would simply modify the definition of $\mathscr{G}$ in 10.24 including all permutations for the iterated integrals and for $\mathscr{K}$ you would use sets of the form $\prod_{i=1}^pA_i$ where A_i is measurable. Everything goes through exactly as above.

Thus the following is mostly obtained.

Theorem 10.14.8 *Let $\{(X_i,\mathscr{F}_i,\mu_i)\}_{i=1}^p$ be σ finite measure spaces and $\prod_{i=1}^p \mathscr{F}_i$ denotes the smallest σ algebra which contains the measurable boxes of the form $\prod_{i=1}^p A_i$ where $A_i \in \mathscr{F}_i$. Then there exists a measure λ defined on $\prod_{i=1}^p \mathscr{F}_i$ such that if $f : \prod_{i=1}^p X_i \to [0,\infty]$ is $\prod_{i=1}^p \mathscr{F}_i$ measurable, $(i_1,\cdots,i_p)$ is any permutation of $(1,\cdots,p)$, then*

$$\int f d\lambda = \int_{X_{i_p}} \cdots \int_{X_{i_1}} f d\mu_{i_1} \cdots d\mu_{i_p} \tag{10.28}$$

If each X_i is a complete separable metric space such that μ_i is finite on balls and $\mathscr{F}_i$ contains $\mathscr{B}(X_i)$, the Borel sets of X_i, then λ is a regular measure on a σ algebra of sets of $\prod_{i=1}^p X_i$ with the metric given by $d(\boldsymbol{x},\boldsymbol{y}) \equiv \max\{d(x_i,y_i) : x_i,y_i \in X_i\}$, which includes the Borel sets.

Proof: It remains to verify the last claim. This is because all sets $\prod_{i=1}^p B(\xi_i,r)$ are contained in $\prod_{i=1}^p \mathscr{F}_i$ and are the open balls for the topology of $\prod_{i=1}^p X_i$. Then by separability of each X_i, the product $\prod_{i=1}^p X_i$ is also separable and so this product with the above metric is completely separable. Thus every open set is the countable union of these sets so open sets are in $\prod_{i=1}^p \mathscr{F}_i$ which consequently contains the Borel sets. Now from Corollary 9.4.9, λ is regular because it is finite on balls. ■

The conclusion 10.28 is called Fubini's theorem. More generally

Theorem 10.14.9 *Suppose, in the situation of Theorem 10.14.8 $f \in L^1$ with respect to the measure λ. Then 10.28 continues to hold.*

Proof: It suffices to prove this for f having real values because if this is shown the general case is obtained by taking real and imaginary parts. Since $f \in L^1\left(\prod_{i=1}^p X_i\right)$, $\int |f| d\lambda < \infty$ and so both $\frac{1}{2}(|f|+f)$ and $\frac{1}{2}(|f|-f)$ are in $L^1\left(\prod_{i=1}^p X_i\right)$ and are each non-negative. Hence from Theorem 10.14.8,

$$\int f d\lambda = \int \left[\frac{1}{2}(|f|+f) - \frac{1}{2}(|f|-f)\right] d\lambda = \int \frac{1}{2}(|f|+f)\, d\lambda - \int \frac{1}{2}(|f|-f)\, d\lambda$$

$$= \int \cdots \int \frac{1}{2}(|f|+f)\, d\mu_{i_1} \cdots d\mu_{i_p} - \int \cdots \int \frac{1}{2}(|f|-f)\, d\mu_{i_1} \cdots d\mu_{i_p}$$

$$= \int \cdots \int \frac{1}{2}(|f|+f) - \frac{1}{2}(|f|-f)\, d\mu_{i_1} \cdots d\mu_{i_p} = \int \cdots \int f d\mu_{i_1} \cdots d\mu_{i_p} \quad ■$$

The following corollary is a convenient way to verify the hypotheses of the above theorem.

Corollary 10.14.10 *Suppose f is measurable with respect to $\prod_{i=1}^p \mathscr{F}_i$ and suppose for some permutation, $(i_1,\cdots,i_p)$, $\int \cdots \int |f| d\mu_{i_1} \cdots d\mu_{i_p} < \infty$. Then $f \in L^1\left(\prod_{i=1}^p X_i\right)$.*

Proof: By Theorem 10.14.8, $\int_{\mathbb{R}^p} |f| d\lambda = \int \cdots \int |f(\boldsymbol{x})| d\mu_{i_1} \cdots d\mu_{i_p} < \infty$ and so f is in $L^1(\mathbb{R}^p)$. ■

You can of course consider the completion of a product measure by using the outer measure approach described earlier. This could have been used to get Lebesgue measure.

If you have $f \geq 0$ and you consider the completion of $\left(\prod_{i=1}^p X_i, \prod_{i=1}^p \mathscr{F}_i, \prod_{i=1}^p \mu_i\right)$ denoted by

$$\left(\prod_{i=1}^p X_i, \overline{\prod_{i=1}^p \mathscr{F}_i}, \overline{\prod_{i=1}^p \mu_i}\right)$$

and f is $\overline{\prod_{i=1}^{p}\mathscr{F}_i}$ measurable, then the procedure for completing a measure space implies there are $h \geq f \geq g$ where h, g are both $\prod_{i=1}^{p}\mathscr{F}_i$ measurable and $f = g = h$ a.e. relative to the complete measure. Here $(X_i, \mathscr{F}_i, \mu_i)$ is finite or σ finite. Then you can define the iterated integral of f as that single number which is between the iterated integrals of g and h both of which make sense. Thus one can make sense of an iterated integral even if the function is not product measurable as long as the function is measurable in the completion of the product measure space. See Problem 19 on Page 348.

10.15 Exercises

1. Let $\Omega = \mathbb{N} = \{1, 2, \cdots\}$. Let $\mathscr{F} = \mathscr{P}(\mathbb{N})$, the set of all subsets of $\mathbb{N}$, and let $\mu(S) =$ number of elements in S. Thus $\mu(\{1\}) = 1 = \mu(\{2\})$, $\mu(\{1,2\}) = 2$, etc. In this case, all functions are measurable. For a nonnegative function, f defined on $\mathbb{N}$, show $\int_{\mathbb{N}} f d\mu = \sum_{k=1}^{\infty} f(k)$. What do the monotone convergence and dominated convergence theorems say about this example?

2. For the measure space of Problem 1, give an example of a sequence of nonnegative measurable functions $\{f_n\}$ converging pointwise to a function f, such that inequality is obtained in Fatou's lemma.

3. If $(\Omega, \mathscr{F}, \mu)$ is a measure space and $f \geq 0$ is measurable, show that if $g(\omega) = f(\omega)$ a.e. ω and $g \geq 0$, then $\int g d\mu = \int f d\mu$. Show that if $f, g \in L^1(\Omega)$ and $g(\omega) = f(\omega)$ a.e. then $\int g d\mu = \int f d\mu$.

4. Let $\{f_n\}, f$ be measurable functions with values in $\mathbb{C}$. $\{f_n\}$ converges in measure if $\lim_{n\to\infty} \mu(x \in \Omega : |f(x) - f_n(x)| \geq \varepsilon) = 0$ for each fixed $\varepsilon > 0$. Prove the theorem of F. Riesz. If f_n converges to f in measure, then there exists a subsequence $\{f_{n_k}\}$ which converges to f a.e. In case μ is a probability measure, this is called convergence in probability. It does not imply pointwise convergence but does imply that there is a subsequence which converges pointwise off a set of measure zero. **Hint:** Choose n_1 such that
$$\mu(x : |f(x) - f_{n_1}(x)| \geq 1) < 1/2.$$
Choose $n_2 > n_1$ such that
$$\mu(x : |f(x) - f_{n_2}(x)| \geq 1/2) < 1/2^2,$$
$n_3 > n_2$ such that
$$\mu(x : |f(x) - f_{n_3}(x)| \geq 1/3) < 1/2^3,$$
etc. Now consider what it means for $f_{n_k}(x)$ to fail to converge to $f(x)$. Use the Borel Cantelli Lemma 9.2.5 on Page 247.

5. Let $(X, \mathscr{F}, \mu)$ be a regular measure space. For example, it could be $\mathbb{R}^p$ with Lebesgue measure. Why do we care about a measure space being regular? This problem will show why. Suppose that closures of balls are compact as in the case of $\mathbb{R}^p$.

   (a) Let $\mu(E) < \infty$. By regularity, there exists $K \subseteq E \subseteq V$ where K is compact and V is open such that $\mu(V \setminus K) < \varepsilon$. Show there exists W open such that $K \subseteq \bar{W} \subseteq V$ and $\bar{W}$ is compact. Now show there exists a function h such that h has values in $[0,1]$, $h(x) = 1$ for $x \in K$, and $h(x)$ equals 0 off W. **Hint:** You might consider Problem 11 on Page 255.

(b) Show that $\int |\mathscr{X}_E - h|\, d\mu < \varepsilon$

(c) Next suppose $s = \sum_{i=1}^n c_i \mathscr{X}_{E_i}$ is a nonnegative simple function where each $\mu(E_i) < \infty$. Show there exists a continuous nonnegative function h which equals zero off some compact set such that $\int |s-h|\, d\mu < \varepsilon$

(d) Now suppose $f \geq 0$ and $f \in L^1(\Omega)$. Show that there exists $h \geq 0$ which is continuous and equals zero off a compact set such that $\int |f-h|\, d\mu < \varepsilon$

(e) If $f \in L^1(\Omega)$ with complex values, show the conclusion in the above part of this problem is the same.

6. Let $(\Omega, \mathscr{F}, \mu)$ be a measure space and suppose $f, g : \Omega \to (-\infty, \infty]$ are measurable. Prove the sets

$$\{\omega : f(\omega) < g(\omega)\} \text{ and } \{\omega : f(\omega) = g(\omega)\}$$

are measurable. **Hint:** The easy way to do this is to write

$$\{\omega : f(\omega) < g(\omega)\} = \cup_{r \in \mathbb{Q}} [f < r] \cap [g > r].$$

Note that $l(x,y) = x - y$ is not continuous on $(-\infty, \infty]$ so the obvious idea doesn't work. Here $[g > r]$ signifies $\{\omega : g(\omega) > r\}$.

7. Let $\{f_n\}$ be a sequence of real or complex valued measurable functions. Let

$$S = \{\omega : \{f_n(\omega)\} \text{ converges}\}.$$

Show S is measurable. **Hint:** You might try to exhibit the set where f_n converges in terms of countable unions and intersections using the definition of a Cauchy sequence.

8. Suppose $u_n(t)$ is a differentiable function for $t \in (a,b)$ and suppose that for $t \in (a,b)$, $|u_n(t)|$, $|u_n'(t)| < K_n$ where $\sum_{n=1}^\infty K_n < \infty$. Show $(\sum_{n=1}^\infty u_n(t))' = \sum_{n=1}^\infty u_n'(t)$.

Hint: This is an exercise in the use of the dominated convergence theorem and the mean value theorem.

9. Suppose $\{f_n\}$ is a sequence of nonnegative measurable functions defined on a measure space, $(\Omega, \mathscr{S}, \mu)$. Show that $\int \sum_{k=1}^\infty f_k d\mu = \sum_{k=1}^\infty \int f_k d\mu$. **Hint:** Use the monotone convergence theorem along with the fact the integral is linear.

10. Explain why for each $t > 0$, $x \to e^{-tx}$ is a function in $L^1(\mathbb{R})$ and $\int_0^\infty e^{-tx} dx = \frac{1}{t}$. Thus

$$\int_0^R \frac{\sin(t)}{t} dt = \int_0^R \int_0^\infty \sin(t)\, e^{-tx} dx dt$$

Now explain why you can change the order of integration in the above iterated integral. Then compute what you get. Next pass to a limit as $R \to \infty$ and show $\int_0^\infty \frac{\sin(t)}{t} dt = \frac{1}{2}\pi$. This is a very important integral. Note that the thing on the left is an improper integral. $\sin(t)/t$ is not Lebesgue integrable because it is not absolutely integrable. That is $\int_0^\infty \left|\frac{\sin t}{t}\right| dm = \infty$. It is important to understand that the Lebesgue theory of integration only applies to nonnegative functions and those which are absolutely integrable.

11. Let the rational numbers in $[0,1]$ be $\{r_k\}_{k=1}^{\infty}$ and define

$$f_n(t) = \begin{cases} 1 \text{ if } t \in \{r_1, \cdots, r_n\} \\ 0 \text{ if } t \notin \{r_1, \cdots, r_n\} \end{cases}$$

Show that $\lim_{n\to\infty} f_n(t) = f(t)$ where f is one on the rational numbers and 0 on the irrational numbers. Explain why each f_n is Riemann integrable but f is not. However, each f_n is actually a simple function and its Lebesgue and Riemann integral is equal to 0. Apply the monotone convergence theorem to conclude that f is Lebesgue integrable and in fact, $\int f dm = 0$.

12. Show $\lim_{n\to\infty} \frac{n}{2^n} \sum_{k=1}^{n} \frac{2^k}{k} = 2$. This problem was shown to me by Shane Tang, a former student. It is a nice exercise in dominated convergence theorem if you massage it a little. **Hint:**

$$\frac{n}{2^n} \sum_{k=1}^{n} \frac{2^k}{k} = \sum_{k=1}^{n} 2^{k-n} \frac{n}{k} = \sum_{l=0}^{n-1} 2^{-l} \frac{n}{n-l} = \sum_{l=0}^{n-1} 2^{-l} \left(1 + \frac{l}{n-l}\right) \leq \sum_{l}^{n-1} 2^{-l} (1+l)$$

13. Suppose you have a real vector space $E(\omega)$ which is a subspace of a normed linear space V, this for each $\omega \in \Omega$ where $(\Omega, \mathscr{F})$ is a measurable space. Suppose $E(\omega) = \text{span}(b_1(\omega), \cdots, b_n(\omega))$ where these $b_i(\omega)$ are linearly independent and each is measurable into V. Define $\theta(\omega) : \mathbb{R}^n \to E(\omega)$ by $\theta(\omega)(\sum_{i=1}^n a_i e_i) \equiv \sum_{i=1}^n a_i b_i(\omega)$. Show that $\theta(\omega)$ maps functions measurable into $\mathbb{R}^n$ to functions measurable into V. Now show $\theta(\omega)^{-1}$ also maps functions measurable into V to functions measurable into $\mathbb{R}^n$. **Hint:** For the second part you need to start with a function $\omega \to h(\omega)$ which is measurable into V with values in $E(\omega)$. Thus $h(\omega) = \sum_{i=1}^n a_i(\omega) b_i(\omega)$. You need to verify that the a_i are measurable. To do this, assume first that $\|h(\omega)\|$ is bounded by some constant M. Then consider $S_r \equiv \left\{\omega : \inf_{|\boldsymbol{a}|>r} \|\sum_i a_i b_i(\omega)\| > M\right\}$. Explain why every ω is in some S_r. Then consider $\Phi(\boldsymbol{a}, \omega)$ which will be defined as $-\|\sum_i a_i b_i(\omega) - h(\omega)\|$ for $\omega \in S_r$. Thus the maximum of this functional for $\omega \in \Omega$ is 0. Show that for $\omega \in S_r$ the maximum will occur on the set $|\boldsymbol{a}| \leq M+1$. Then apply Kuratowski's lemma. Finally consider a truncation of h called h_m and apply what was just shown to this truncation which has $\|h_m(\omega)\| \leq m$. Then let $m \to \infty$ and observe that for large enough $m, h_m(\omega) = h(\omega)$ and so the $a_i^m(\omega)$ are also constant from this value onward. Thus $\lim_{m\to\infty} a_i^m(\omega) \equiv a_i(\omega)$ exists and a_i is measurable, being the limit of measurable functions.

14. Give an example of a sequence of functions $\{f_n\}, f_n \geq 0$ and a function $f \geq 0$ such that $f(x) = \liminf_{n\to\infty} f_n(x)$ but $\int f dm < \liminf_{n\to\infty} \int f_n dm$ so you get strict inequality in Fatou's lemma.

15. Let f be a nonnegative Riemann integrable function defined on $[a,b]$. Thus there is a unique number between all the upper sums and lower sums. First explain why, if $a_i \geq 0, \int \sum_{i=1}^n a_i \mathscr{X}_{[t_i, t_{i-1})}(t)\, dm = \sum_i a_i (t_i - t_{i-1})$. Explain why there exists an increasing sequence of Borel measurable functions $\{g_n\}$ converging to a Borel measurable function g, and a decreasing sequence of functions $\{h_n\}$ which are also Borel measurable converging to a Borel measurable function h such that $g_n \leq f \leq h_n$,

$$\int g_n dm \text{ equals a lower sum, } \int h_n dm \text{ equals an upper sum}$$

and $\int (h-g)\,dm = 0$. Explain why $\{x : f(x) \neq g(x)\}$ is a set of measure zero. Then explain why f is measurable and $\int_a^b f(x)\,dx = \int f dm$ so that the Riemann integral gives the same answer as the Lebesgue integral. Here m is one dimensional Lebesgue measure discussed earlier.

16. Let $(\Omega, \mathscr{F}, \mu)$ be a measure space. Suppose f, g are nonnegative and measurable functions. Show that $\int_\Omega f g d\mu = \int_0^\infty \int_{[g>y]} f d\mu dm$. This is due to Faddeyev. **Hint:** It is a little involved. First suppose g is a simple function. Then approximate arbitrary g with an increasing sequence of such functions and pass to a limit using the monotone convergence theorem. Here $[g > y] \equiv \{\omega : g(\omega) > y\}$.

17. Let λ, μ be finite measures. We say $\lambda \ll \mu$ if whenever $\mu(E) = 0$ it follows $\lambda(E) = 0$. Show that if $\lambda \ll \mu$, then for every $\varepsilon > 0$ there exists $\delta > 0$ such that if $\mu(E) < \delta$, then $\lambda(E) < \varepsilon$.

18. If λ is a signed measure with values in $\mathbb{R}$ so that when $\{E_i\}$ are disjoint, $\sum_i \lambda(E_i)$ converges, show that the infinite series converges absolutely also.

19. In the Radon Nikodym Theorem 10.13.7, show that if $f, \hat{f}$ both work, then $f = \hat{f}$ a.e.

20. Suppose $\nu \ll \mu$ where these are finite measures so there exists $h \geq 0$ and measurable such that $\nu(E) = \int_E h d\mu$ by the Radon Nikodym theorem. Show that if f is measurable and non-negative, then $\int f d\nu = \int f h d\mu$. **Hint:** It holds if f is χ_E and so it holds for a simple function. Now consider a sequence of simple functions increasing to f and use the monotone convergence theorem.

Chapter 11

Regular Measures

So far, most examples have been in one dimensional settings. This is about to change.

11.1 Regular Measures in a Metric Space

In this section X will be a metric space in which the closed balls are compact. The extra generality involving a metric space instead of $\mathbb{R}^p$ would allow the consideration of manifolds for example. However, $\mathbb{R}^p$ is an important case.

Definition 11.1.1 *The symbol $C_c(V)$ for V an open set will denote the continuous functions having compact support which is contained in V. Recall that the support of a continuous function f is defined as the closure of the set on which the function is nonzero. $L : C_c(X) \to \mathbb{C}$ is called a positive linear functional if it is linear, $L(\alpha f + \beta g) = \alpha Lf + \beta Lg$ and satisfies $Lf \leq Lg$ if $f \leq g$. Also, recall that a measure μ is regular on some σ algebra $\mathcal{F}$ containing the Borel sets if for every $F \in \mathcal{F}$,*

$$\begin{aligned} \mu(F) &= \sup\{\mu(K) : K \subseteq F \text{ and } K \text{ compact}\} \\ \mu(F) &= \inf\{\mu(V) : V \supseteq F \text{ and } V \text{ is open}\} \end{aligned}$$

A complete measure, finite on compact sets, which is regular as above, is called a Radon measure. A set is called an F_σ set if it is the countable union of closed sets and a set is G_δ if it is the countable intersection of open sets.

Remarkable things happen in the above context. Some are described in the following proposition.

Proposition 11.1.2 *Suppose (X,d) is a metric space in which the closed balls are compact and X is a countable union of closed balls. Also suppose $(X,\mathcal{F},\mu)$ is a complete measure space, $\mathcal{F}$ contains the Borel sets, and that μ is regular and finite on measurable subsets of finite balls. Then*

1. *For each $E \in \mathcal{F}$, there is an F_σ set F and a G_δ set G such that $F \subseteq E \subseteq G$ and $\mu(G \setminus F) = 0$.*

2. *Also if $f \geq 0$ is $\mathcal{F}$ measurable, then there exists $g \leq f$ such that g is Borel measurable and $g = f$ a.e.*

   and $h \geq f$ such that h is Borel measurable and $h = f$ a.e.

3. *If $E \in \mathcal{F}$ is a bounded set contained in a ball $B(x_0,r) = V$, then there exists a sequence of continuous functions in $C_c(V)$ $\{h_n\}$ having values in $[0,1]$ and a set of measure zero N such that for $x \notin N$, $h_n(x) \to \mathcal{X}_E(x)$. Also $\int |h_n - \mathcal{X}_E| \, d\mu \to 0$. Letting $\tilde{N}$ be a G_δ set of measure zero containing N, $h_n \mathcal{X}_{\tilde{N}^C} \to \mathcal{X}_F$ where $F \subseteq E$ and $\mu(E \setminus F) = 0$.*

4. *If $f \in L^1(X,\mathcal{F},\mu)$, there exists $g \in C_c(X)$, such that $\int_X |f - g| \, d\mu < \varepsilon$. There also exists a sequence of functions in $C_c(X)$ $\{g_n\}$ which converges pointwise to f.*

Proof: 1. Let $R_n \equiv B(x_0,n), R_0 = \emptyset$. If E is measurable, let $E_n \equiv E \cap (R_n \setminus R_{n-1})$. Thus these E_n are disjoint and their union is E. By outer regularity, there exists open $U_n \supseteq E_n$ such that $\mu(U_n \setminus E_n) < \varepsilon/2^n$. Now if $U \equiv \cup_n U_n$, it follows that $U \setminus E \subseteq \cup_n (U_n \setminus E_n)$ so $\mu(U \setminus E) \leq \sum_{n=1}^{\infty} \frac{\varepsilon}{2^n} = \varepsilon$. This has shown that there exists an open set U containing E such that $\mu(U \setminus E) \leq \varepsilon$. Let V_n be open, containing E and $\mu(V_n \setminus E) < \frac{1}{2^n}, V_n \supseteq V_{n+1}$. Let $G \equiv \cap_n V_n$. This is a G_δ set containing E and $\mu(G \setminus E) \leq \mu(V_n \setminus E) < \frac{1}{2^n}$ and so $\mu(G \setminus E) = 0$. By inner regularity, there is F_n an F_σ set contained in E_n with $\mu(E_n \setminus F_n) = 0$. Then let $F \equiv \cup_n F_n$. This F is an F_σ set and $\mu(E \setminus F) \leq \sum_n \mu(E_n \setminus F_n) = 0$. Thus $F \subseteq E \subseteq G$ and $\mu(G \setminus F) \leq \mu(G \setminus E) + \mu(E \setminus F) = 0$.

2. If f is measurable and nonnegative, from Theorem 9.1.6 there is an increasing sequence of simple functions s_n such that $\lim_{n\to\infty} s_n(x) = f(x)$. Say $s_n(x) \equiv \sum_{k=1}^{m_n} c_k^n \mathcal{X}_{E_k^n}(x)$. Let $m_p\left(E_k^n \setminus F_k^n\right) = 0$ where F_k^n is an F_σ set. Replace E_k^n with F_k^n and let $\tilde{s}_n$ be the resulting simple function. Let $g(x) \equiv \lim_{n\to\infty} \tilde{s}_n(x)$. Then g is Borel measurable and $g \leq f$ and $g = f$ except for a set of measure zero, the union of the sets where s_n is not equal to $\tilde{s}_n$. As to the other claim, let $h_n(x) \equiv \sum_{k=1}^{\infty} \mathcal{X}_{A_{kn}}(x) \frac{k}{2^n}$ where A_{kn} is a G_δ set containing $f^{-1}\left(\left(\frac{k-1}{2^n}, \frac{k}{2^n}\right]\right)$ for which $\mu\left(A_{kn} \setminus f^{-1}\left(\left(\frac{k-1}{2^n}, \frac{k}{2^n}\right]\right)\right) \equiv \mu(D_{kn}) = 0$. If $N = \cup_{k,n} D_{kn}$, then N is a set of measure zero. On N^C, $h_n(x) \to f(x)$. Let $h(x) = \liminf_{n\to\infty} h_n(x)$. Note that $\mathcal{X}_{A_{kn}}(x) \frac{k}{2^n} \geq \mathcal{X}_{f^{-1}\left(\left(\frac{k-1}{2^n}, \frac{k}{2^n}\right]\right)}(x) \frac{k}{2^n}$ and so $h_n(x) \geq h(x)$ and $\liminf_{n\to\infty} h_n(x)$ is Borel measurable because each h_n is.

3. Let $K_n \subseteq E \subseteq V_n$ with K_n compact and V_n open such that $V_n \subseteq B(x_0,r)$ and

$$\mu(V_n \setminus K_n) < 2^{-(n+1)}.$$

Then from Lemma 3.12.4, there exists h_n with $K_n \prec h_n \prec V_n$. Then $\int |h_n - \mathcal{X}_E| \, d\mu < 2^{-n}$ and so

$$\mu\left(|h_n - \mathcal{X}_E| > \left(\frac{2}{3}\right)^n\right) < \left(\left(\frac{3}{2}\right)^n \int_{\left[|h_n - \mathcal{X}_E| > \left(\frac{2}{3}\right)^n\right]} |h_n - \mathcal{X}_E| \, d\mu\right) \leq \left(\frac{3}{4}\right)^n$$

By Lemma 9.2.5 there is a set of measure zero N such that if $x \notin N$, it is in only finitely many of the sets $\left[|h_n - \mathcal{X}_E| > \left(\frac{2}{3}\right)^n\right]$. Thus on N^C, eventually, for all k large enough, $|h_k - \mathcal{X}_E| \leq \left(\frac{2}{3}\right)^k$ so $h_k(x) \to \mathcal{X}_E(x)$ off N. The assertion about convergence of the integrals follows from the dominated convergence theorem and the fact that each h_n is nonnegative, bounded by 1, $(K_n \prec h_n \prec V_n)$ and is 0 off some ball. In the last claim, it only remains to verify that $h_n \mathcal{X}_{\tilde{N}^C}$ converges to an indicator function because each $h_n \mathcal{X}_{\tilde{N}^C}$ is Borel measurable. ($\tilde{N} \supseteq N$ and $\tilde{N}$ is a Borel set and $\mu\left(\tilde{N} \setminus N\right) = 0$) Thus its limit will also be Borel measurable. However, $h_n \mathcal{X}_{\tilde{N}^C}$ converges to 1 on $E \cap \tilde{N}^C$, 0 on $E^C \cap \tilde{N}^C$ and 0 on $\tilde{N}$. Thus $E \cap \tilde{N}^C = F$ and $h_n \mathcal{X}_{\tilde{N}^C}(x) \to \mathcal{X}_F$ where $F \subseteq E$ and $\mu(E \setminus F) \leq \mu\left(\tilde{N}\right) = 0$.

4. It suffices to assume $f \geq 0$ because you can consider the positive and negative parts of the real and imaginary parts of f and reduce to this case. Let $f_n(x) \equiv \mathcal{X}_{B(x_0,n)}(x) f(x)$. Then by the dominated convergence theorem, if n is large enough, $\int |f - f_n| \, d\mu < \varepsilon$. There is a nonnegative simple function $s \leq f_n$ such that $\int |f_n - s| \, d\mu < \varepsilon$. This follows from picking k large enough in an increasing sequence of simple functions $\{s_k\}$ converging to f_n and the dominated convergence theorem. Say $s(x) = \sum_{k=1}^{m} c_k \mathcal{X}_{E_k}(x)$. Then let $K_k \subseteq E_k \subseteq V_k$ where K_k, V_k are compact and open respectively and $\sum_{k=1}^{m} c_k \mu(V_k \setminus K_k) < \varepsilon$. By Lemma

3.12.4, there exists h_k with $K_k \prec h_k \prec V_k$. Then

$$\begin{aligned}\int \left| \sum_{k=1}^{m} c_k \mathcal{X}_{E_k}(x) - \sum_{k=1}^{m} c_k h_k(x) \right| d\mu &\leq \sum_k c_k \int \left| \mathcal{X}_{E_k}(x) - h_k(x) \right| dx \\ &< 2\sum_k c_k \mu(V_k \setminus K_k) < 2\varepsilon\end{aligned}$$

Let $g \equiv \sum_{k=1}^{m} c_k h_k(x)$. Thus $\int |s-g|\, d\mu \leq 2\varepsilon$. Then

$$\int |f-g|\, d\mu \leq \int |f-f_n|\, d\mu + \int |f_n - s|\, d\mu + \int |s-g|\, d\mu < 4\varepsilon$$

Since ε is arbitrary, this proves the first part of 4. For the second part, let $g_n \in C_c(X)$ such that $\int |f-g_n|\, d\mu < 2^{-n}$. Let $A_n \equiv \left\{ x : |f-g_n| > \left(\frac{2}{3}\right)^n \right\}$. Then

$$\mu(A_n) \leq \left(\frac{3}{2}\right)^n \int_{A_n} |f-g_n|\, d\mu \leq \left(\frac{3}{4}\right)^n.$$

Thus, if N is all x in infinitely many A_n, then by the Borel Cantelli lemma, $\mu(N) = 0$ and if $x \notin N$, then x is in only finitely many A_n and so for all n large enough, $|f(x) - g_n(x)| \leq \left(\frac{2}{3}\right)^n$. ■

11.2 Differentiation of Radon Measures

For the rest of this chapter, I will specialize to $\mathbb{R}^p$ or at least a finite dimensional normed linear space. For different proofs and some results which are not discussed here, a good source is [17] which is where I first read some of these things.

Recall the following Besicovitch covering theorem for Radon measures. It is Corollary 9.12.3 on Page 269 and the earlier version, Theorem 4.5.8 on Page 119 which are listed here for the sake of convenience.

Corollary 11.2.1 *Let E be a nonempty set and let μ be a Radon measure on a σ algebra which contains the Borel sets of $\mathbb{R}^p$. Suppose $\mathcal{F}$ is a collection of closed balls which cover E in the sense of Vitali. Then there exists a sequence of disjoint closed balls $\{B_i\} \subseteq \mathcal{F}$ such that*

$$\overline{\mu}\left(E \setminus \cup_{j=1}^{N} B_j\right) = 0, N \leq \infty$$

Theorem 11.2.2 *There exists a constant N_p, depending only on p with the following property. If $\mathcal{F}$ is any collection of nonempty balls in $\mathbb{R}^p$ with*

$$\sup\{\operatorname{diam}(B) : B \in \mathcal{F}\} = D < \infty$$

and if A is the set of centers of the balls in $\mathcal{F}$, then there exist subsets of $\mathcal{F}$, $\mathcal{H}_1, \cdots, \mathcal{H}_{N_p}$, such that each $\mathcal{H}_i$ is a countable collection of disjoint balls from $\mathcal{F}$ (possibly empty) and

$$A \subseteq \cup_{i=1}^{N_p} \cup \{B : B \in \mathcal{H}_i\}.$$

11.3 Maximal Functions

In this section the Besicovitch covering theorem, Theorem 4.5.8 will be used t the Lebesgue differentiation theorem for general Radon measures. This will end up including Lebesgue measure presented later. In what follows, μ will be a Radon measure, complete, inner and outer regular, and finite on compact sets. Also

$$Z \equiv \{\boldsymbol{x} \in \mathbb{R}^p : \mu(B(\boldsymbol{x},r)) = 0 \text{ for some } r > 0\}, \tag{11.1}$$

Lemma 11.3.1 *Z is measurable and $\mu(Z) = 0$.*

Proof: For each $\boldsymbol{x} \in Z$, there exists a ball $B(\boldsymbol{x},r)$ with $\mu(B(\boldsymbol{x},r)) = 0$. Let $\mathscr{C}$ be the collection of these balls. Since $\mathbb{R}^p$ has a countable basis, a countable subset $\widetilde{\mathscr{C}} \equiv \{B_i\}_{i=1}^{\infty}$, of $\mathscr{C}$ also covers Z. Then letting $\overline{\mu}$ denote the outer measure determined by μ, $\overline{\mu}(Z) \leq \sum_{i=1}^{\infty} \overline{\mu}(B_i) = \sum_{i=1}^{\infty} \mu(B_i) = 0$. Therefore, Z is measurable, $(\bar{\mu}(S) \geq \bar{\mu}(S \cap Z) + \bar{\mu}(S \cap Z^C))$ and has measure zero as claimed. ■

Let $Mf : \mathbb{R}^p \to [0,\infty]$ by

$$Mf(\boldsymbol{x}) \equiv \begin{cases} \sup_{r \leq 1} \frac{1}{\mu(B(\boldsymbol{x},r))} \int_{B(\boldsymbol{x},r)} |f| \, d\mu \text{ if } \boldsymbol{x} \notin Z \\ 0 \text{ if } \boldsymbol{x} \in Z \end{cases}.$$

I will begin using $\|f\|_1$ for the integral $\int_\Omega |f| \, d\mu$.

The special points described in the following theorem are called Lebesgue points.

Theorem 11.3.2 *Let μ be a Radon measure and let $f \in L^1(\mathbb{R}^p, \mu)$ meaning that $\int_\Omega |f| \, d\mu < \infty$. Then for μ a.e. $\boldsymbol{x}$,*

$$\lim_{r \to 0} \frac{1}{\mu(B(\boldsymbol{x},r))} \int_{B(\boldsymbol{x},r)} |f(\boldsymbol{y}) - f(\boldsymbol{x})| \, d\mu(y) = 0$$

Also $\overline{\mu}([Mf > \varepsilon]) \leq N_p \varepsilon^{-1} \|f\|_1$.

Proof: First consider the following claim which is called a weak type estimate.

Claim 1: The following inequality holds for N_p the constant of the Besicovitch covering theorem, Theorem 4.5.8.

$$\overline{\mu}([Mf > \varepsilon]) \leq N_p \varepsilon^{-1} \|f\|_1$$

Proof of claim: First note $[Mf > \varepsilon] \cap Z = \emptyset$ and without loss of generality, you can assume $\overline{\mu}([Mf > \varepsilon]) > 0$. Next, for each $\boldsymbol{x} \in [Mf > \varepsilon]$ there exists a ball $B_{\boldsymbol{x}} = B(\boldsymbol{x}, r_{\boldsymbol{x}})$ with $r_{\boldsymbol{x}} \leq 1$ and

$$\mu(B_{\boldsymbol{x}})^{-1} \int_{B(\boldsymbol{x},r_{\boldsymbol{x}})} |f| \, d\mu > \varepsilon.$$

Let $\mathscr{F}$ be this collection of balls so that $[Mf > \varepsilon]$ is the set of centers of balls of $\mathscr{F}$. By the Besicovitch covering theorem, Theorem 4.5.8,

$$[Mf > \varepsilon] \subseteq \cup_{i=1}^{N_p} \{B : B \in \mathscr{G}_i\}$$

where $\mathscr{G}_i$ is a collection of disjoint balls of $\mathscr{F}$. Now for some i,

$$\overline{\mu}([Mf > \varepsilon]) / N_p \leq \mu(\cup \{B : B \in \mathscr{G}_i\})$$

because if this is not so, then for all $i, \overline{\mu}([Mf>\varepsilon])/N_p > \mu(\cup\{B : B\in\mathscr{G}_i\})$ and so

$$\overline{\mu}([Mf>\varepsilon]) \le \sum_{i=1}^{N_p}\mu(\cup\{B : B\in\mathscr{G}_i\}) < \sum_{i=1}^{N_p}\frac{\overline{\mu}([Mf>\varepsilon])}{N_p} = \overline{\mu}([Mf>\varepsilon]),$$

a contradiction. Therefore for this i,

$$\begin{aligned}\frac{\overline{\mu}([Mf>\varepsilon])}{N_p} &\le \mu(\cup\{B : B\in\mathscr{G}_i\}) = \sum_{B\in\mathscr{G}_i}\mu(B) \le \sum_{B\in\mathscr{G}_i}\varepsilon^{-1}\int_B |f|\,d\mu\\ &\le \varepsilon^{-1}\int_{\mathbb{R}^p}|f|\,d\mu = \varepsilon^{-1}\|f\|_1.\end{aligned}$$

This shows Claim 1.

Claim 2: If g is any continuous function defined on $\mathbb{R}^p$, then for $\boldsymbol{x}\notin Z$,

$$\lim_{r\to 0}\frac{1}{\mu(B(\boldsymbol{x},r))}\int_{B(\boldsymbol{x},r)}|g(\boldsymbol{y})-g(\boldsymbol{x})|\,d\mu(y) = 0$$

and

$$\lim_{r\to 0}\frac{1}{\mu(B(\boldsymbol{x},r))}\int_{B(\boldsymbol{x},r)} g(\boldsymbol{y})\,d\mu(y) = g(\boldsymbol{x}). \tag{11.2}$$

Proof: Since g is continuous at $\boldsymbol{x}$, whenever r is small enough,

$$\frac{1}{\mu(B(\boldsymbol{x},r))}\int_{B(\boldsymbol{x},r)}|g(\boldsymbol{y})-g(\boldsymbol{x})|\,d\mu(y) \le \frac{1}{\mu(B(\boldsymbol{x},r))}\int_{B(\boldsymbol{x},r)}\varepsilon\,d\mu(y) = \varepsilon.$$

11.2 follows from the above and the triangle inequality. This proves the claim.

Now let $g\in C_c(\mathbb{R}^p)$ and $\boldsymbol{x}\notin Z$. Then from the above observations about continuous functions,

$$\overline{\mu}\left(\left[\boldsymbol{x}\notin Z : \limsup_{r\to 0}\frac{1}{\mu(B(\boldsymbol{x},r))}\int_{B(\boldsymbol{x},r)}|f(\boldsymbol{y})-f(\boldsymbol{x})|\,d\mu(y) > \varepsilon\right]\right) \tag{11.3}$$

$$\begin{aligned}&\le \overline{\mu}\left(\left[\boldsymbol{x}\notin Z : \limsup_{r\to 0}\frac{1}{\mu(B(\boldsymbol{x},r))}\int_{B(\boldsymbol{x},r)}|f(\boldsymbol{y})-g(\boldsymbol{y})|\,d\mu(y) > \frac{\varepsilon}{2}\right]\right)\\ &\quad +\overline{\mu}\left(\left[\boldsymbol{x}\notin Z : |g(\boldsymbol{x})-f(\boldsymbol{x})| > \frac{\varepsilon}{2}\right]\right).\\ &\le \overline{\mu}\left(\left[M(f-g) > \frac{\varepsilon}{2}\right]\right) + \overline{\mu}\left(\left[|f-g| > \frac{\varepsilon}{2}\right]\right)\end{aligned} \tag{11.4}$$

Now $\int_{\left[|f-g|>\frac{\varepsilon}{2}\right]}|f-g|\,d\mu \ge \frac{\varepsilon}{2}\overline{\mu}\left(\left[|f-g|>\frac{\varepsilon}{2}\right]\right)$ and so using Claim 1 in 11.4, it follows that 11.3 is dominated by $\left(\frac{2}{\varepsilon}+\frac{N_p}{\varepsilon}\right)\int|f-g|\,d\mu$. But by Theorem 10.8.7, g can be chosen to make this as small as desired. Hence 11.3 is 0. Now observe that

$$\begin{aligned}&\overline{\mu}\left(\left[\boldsymbol{x}\notin Z : \limsup_{r\to 0}\frac{1}{\mu(B(\boldsymbol{x},r))}\int_{B(\boldsymbol{x},r)}|f(\boldsymbol{y})-f(\boldsymbol{x})|\,d\mu(y) > 0\right]\right)\\ \le &\sum_{k=1}^{\infty}\overline{\mu}\left(\left[\boldsymbol{x}\notin Z : \limsup_{r\to 0}\frac{1}{\mu(B(\boldsymbol{x},r))}\int_{B(\boldsymbol{x},r)}|f(\boldsymbol{y})-f(\boldsymbol{x})|\,d\mu(y) > \frac{1}{k}\right]\right) = 0\end{aligned}$$

By completeness of μ this implies

$$\left[\boldsymbol{x}\notin Z\ :\limsup_{r\to 0}\frac{1}{\mu\left(B\left(\boldsymbol{x},r\right)\right)}\int_{B(\boldsymbol{x},r)}\left|f\left(\boldsymbol{y}\right)-f\left(\boldsymbol{x}\right)\right|d\mu\left(y\right)>0\right]$$

is a set of μ measure zero. ■

The following corollary is the main result referred to as the Lebesgue Besicovitch Differentiation theorem.

Definition 11.3.3 *$f\in L^1_{loc}\left(\mathbb{R}^p,\mu\right)$ means $f\mathcal{X}_B$ is in $L^1\left(\mathbb{R}^p,\mu\right)$ whenever B is a ball.*

Corollary 11.3.4 *If $f\in L^1_{loc}\left(\mathbb{R}^p,\mu\right)$, then for μ a.e.$\boldsymbol{x}\notin Z$,*

$$\lim_{r\to 0}\frac{1}{\mu\left(B\left(\boldsymbol{x},r\right)\right)}\int_{B(\boldsymbol{x},r)}\left|f\left(\boldsymbol{y}\right)-f\left(\boldsymbol{x}\right)\right|d\mu\left(y\right)=0\,. \tag{11.5}$$

Proof: If f is replaced by $f\mathcal{X}_{B(\mathbf{0},k)}$ then the conclusion 11.5 holds for all $\boldsymbol{x}\notin F_k$ where F_k is a set of μ measure 0. Letting $k=1,2,\cdots$, and $F\equiv\cup_{k=1}^{\infty}F_k$, it follows that F is a set of measure zero and for any $\boldsymbol{x}\notin F$, and $k\in\{1,2,\cdots\}$, 11.5 holds if f is replaced by $f\mathcal{X}_{B(\mathbf{0},k)}$. Picking any such $\boldsymbol{x}$, and letting $k>|\boldsymbol{x}|+1$, this shows

$$\begin{aligned}&\lim_{r\to 0}\frac{1}{\mu\left(B\left(\boldsymbol{x},r\right)\right)}\int_{B(\boldsymbol{x},r)}\left|f\left(\boldsymbol{y}\right)-f\left(\boldsymbol{x}\right)\right|d\mu\left(y\right)\\ =\ &\lim_{r\to 0}\frac{1}{\mu\left(B\left(\boldsymbol{x},r\right)\right)}\int_{B(\boldsymbol{x},r)}\left|f\mathcal{X}_{B(\mathbf{0},k)}\left(\boldsymbol{y}\right)-f\mathcal{X}_{B(\mathbf{0},k)}\left(\boldsymbol{x}\right)\right|d\mu\left(y\right)=0\end{aligned}$$

because for all r small enough, $B\left(\boldsymbol{x},r\right)\subseteq B\left(\mathbf{0},k\right)$. ■

Definition 11.3.5 *Let E be a measurable set. Then $\boldsymbol{x}\in E$ is called a point of density if $\boldsymbol{x}\notin Z$ and $\lim_{r\to 0}\frac{\mu(B(\boldsymbol{x},r)\cap E)}{\mu(B(\boldsymbol{x},r))}=1$. Recall Z is the set of points $\boldsymbol{x}$ where $\mu\left(B\left(\boldsymbol{x},r\right)\right)=0$ for some $r>0$.*

Proposition 11.3.6 *Let E be a measurable set. Then μ a.e. $\boldsymbol{x}\in E$ is a point of density.*

Proof: This follows from letting $f\left(\boldsymbol{x}\right)=\mathcal{X}_E\left(\boldsymbol{x}\right)$ in Corollary 11.3.4. ■

11.4 Strong Estimates for Maximal Function

From Theorem 11.3.2, $\mu\left(Mf>\lambda\right)\leq\frac{C}{\lambda}\left\|f\right\|_{L^1}$ where C depends on the dimension only. I will denote by $\left\|f\right\|_{L^p}$ the expression $\left(\int\left|f\left(\boldsymbol{x}\right)\right|^p d\mu\right)^{1/p}$to coincide with later notation.

$$\begin{aligned}Mf\ &\equiv\ \sup_{r\leq 1}\frac{1}{\mu\left(B\left(\boldsymbol{x},r\right)\right)}\int_{B(\boldsymbol{x},r)}\left|f\right|d\mu\leq\sup_{r\leq 1}\frac{1}{\mu\left(B\left(\boldsymbol{x},r\right)\right)}\int_{B(\boldsymbol{x},r)}\left|f\right|\mathcal{X}_{\left[|f|>\frac{\lambda}{2}\right]}+\frac{\lambda}{2}d\mu\\ &=\ M\left(f\mathcal{X}_{\left[|f|>\frac{\lambda}{2}\right]}\right)+\frac{\lambda}{2}\end{aligned}$$

Therefore, $\left[Mf>\lambda\right]\subseteq\left[M\left(f\mathcal{X}_{\left[|f|>\frac{\lambda}{2}\right]}\right)>\lambda/2\right]$. It follows that

$$\mu\left(\left[Mf>\lambda\right]\right)\leq\frac{C}{\lambda}\int\left|f\right|\mathcal{X}_{\left[|f|>\lambda/2\right]}$$

$$\int |Mf|^p \, d\mu = \int_0^\infty \mu\left(\left[Mf > \lambda^{1/p}\right]\right) d\lambda \leq \int_0^\infty \frac{C}{\lambda^{1/p}} \int_{\left[|f| > \frac{\lambda^{1/p}}{2}\right]} |f| \, d\mu d\lambda$$

$$= C\int_0^\infty \frac{1}{\eta} \int_{\left[|f| > \frac{\eta}{2}\right]} |f| \, d\mu p\eta^{p-1} d\eta = Cp \int |f(\boldsymbol{x})| \int_0^{2|f|} \eta^{p-2} d\eta d\mu$$

$$= Cp \int |f(\boldsymbol{x})| \frac{2^{p-1}}{p-1} |f(\boldsymbol{x})|^{p-1} d\mu = \frac{Cp2^{p-1}}{p-1} \|f\|_{L^p}^p$$

Of course this is all assuming that Mf is measurable. This is most easily shown if $\mu = m$ Lebesgue measure and this is the case of most interest to me. Consider

$$\boldsymbol{x} \to \int_{B(\boldsymbol{x},r)} |f(\boldsymbol{y})| \, dm \equiv f_r(\boldsymbol{x})$$

This is continuous assuming $f \in L^1_{loc}$. Thus $\boldsymbol{x} \to \frac{1}{B(\boldsymbol{x},r)} \int_{B(\boldsymbol{x},r)} |f(\boldsymbol{y})| \, dm$ is continuous. Now it follows that $Mf(\boldsymbol{x}) = \sup_{0<r<1} f_r(\boldsymbol{x}) = \sup_{0<r<1, r\in\mathbb{Q}} f_r(\boldsymbol{x})$ is Borel measurable. Therefore, we can state the following corollary.

Corollary 11.4.1 *Let $\|f\|_{L^p} < \infty$. Then $\|Mf\|_{L^p} \leq C\|f\|_{L^p}$ where C depends only on $p > 1$ and the dimension. This is called a strong estimate for the maximal function as opposed to the one for $p = 1$ which is called a weak estimate.*

11.5 Symmetric Derivative for Radon Measures

Here we have two Radon measures μ, λ defined on a σ algebra of sets $\mathscr{F}$ which are subsets of an open subset U of $\mathbb{R}^p$, possibly all of $\mathbb{R}^p$. They are complete and Borel and inner and outer regular, and finite on compact sets. Thus both of these measures are σ finite.

In this section is the symmetric derivative $D_\mu(\lambda)$. In what follows, $B(\boldsymbol{x},r)$ will denote a **closed** ball with center $\boldsymbol{x}$ and radius r. Also, let λ and μ be Radon measures and as above, Z will denote a μ measure zero set off of which $\mu(B(\boldsymbol{x},r)) > 0$ for all $r > 0$. Generalizing the notion of lim sup and lim inf,

$$\limsup_{r\to 0} f(r) \equiv \lim_{r\to 0}\left(\sup\{f(t) : t < r\}\right), \ \liminf_{r\to 0} f(r) \equiv \lim_{r\to 0}\left(\inf\{f(t) : t < r\}\right)$$

Then directly from this definition, the $\lim_{r\to 0}$ exists if and only if these two are equal.

Definition 11.5.1 *For $\boldsymbol{x} \notin Z$, define the upper and lower symmetric derivatives as*

$$\overline{D}_\mu\lambda(\boldsymbol{x}) \equiv \limsup_{r\to 0} \frac{\lambda(B(\boldsymbol{x},r))}{\mu(B(\boldsymbol{x},r))}, \ \underline{D}_\mu\lambda(\boldsymbol{x}) \equiv \liminf_{r\to 0} \frac{\lambda(B(\boldsymbol{x},r))}{\mu(B(\boldsymbol{x},r))}.$$

respectively. Also define $D_\mu\lambda(\boldsymbol{x}) \equiv \overline{D}_\mu\lambda(\boldsymbol{x}) = \underline{D}_\mu\lambda(\boldsymbol{x})$ in the case when both the upper and lower derivatives are equal. Recall that $Z \equiv \{\boldsymbol{x} : \mu(B(\boldsymbol{x},r)) = 0 \text{ for some } r > 0\}$ and that this set has measure zero.

Lemma 11.5.2 *Let λ and μ be Radon measures on $\mathscr{F}_\lambda$ and $\mathscr{F}_\mu$ respectively and let $a, b > 0$. If A is a subset of $\left\{\boldsymbol{x} \notin Z : \overline{D}_\mu\lambda(\boldsymbol{x}) \geq b\right\}$ then $\overline{\lambda}(A) \geq b\overline{\mu}(A)$ and if A is a subset of $\left\{\boldsymbol{x} \notin Z : \underline{D}_\mu\lambda(\boldsymbol{x}) \leq a\right\}$, then $\overline{\lambda}(A) \leq a\overline{\mu}(A)$.*

Proof: Let $\overline{\lambda}$ be the outer measure determined by λ, similar for $\overline{\mu}$ and μ. Suppose first that A is a subset of $\{x \notin Z : \overline{D}_\mu \lambda(x) \geq b\}$ so $\mu(B(x,r)) > 0$ for all $r > 0$ and $\overline{\lambda}(A) < \infty$. Let small $\varepsilon > 0$, and let V be a bounded open set with $V \supseteq A$ and $\lambda(V) - \varepsilon < \overline{\lambda}(A)$. Then for each $x \in A$,

$$\frac{\lambda(B(x,r))}{\mu(B(x,r))} > b - \varepsilon,\ B(x,r) \subseteq V,$$

for infinitely many values of r which are arbitrarily small. Thus the collection of such closed balls constitutes a Vitali cover for A. By Corollary 9.12.3 there is a disjoint sequence of these closed balls $\{B_i\}$ such that

$$\overline{\mu}(A \setminus \cup_{i=1}^{\infty} B_i) = 0,$$

$$\overline{\mu}(A) \leq \overline{\mu}(A \setminus \cup_{i=1}^{\infty} B_i) + \overline{\mu}(\cup_{i=1}^{\infty} B_i \cap A) \leq \sum_{i=1}^{\infty} \overline{\mu}(B_i \cap A) \tag{11.6}$$

$$\leq \sum_{i=1}^{\infty} \mu(B_i) < \frac{1}{b-\varepsilon} \sum_{i=1}^{\infty} \lambda(B_i) \leq \frac{1}{b-\varepsilon} \lambda(V) < \frac{\varepsilon + \overline{\lambda}(A)}{b-\varepsilon}$$

Since ε is arbitrary, this shows $\overline{\mu}(A) b \leq \overline{\lambda}(A)$. In case $\overline{\lambda}(A) = \infty$, there is nothing to show.

Now suppose A is a subset of $\{x \notin Z : \underline{D}_\mu \lambda(x) \leq a\}$. Suppose first that A is bounded so $\overline{\mu}(A) < \infty$ and let V be a bounded open set containing A with $\mu(V) - \varepsilon < \overline{\mu}(A)$. Then for each $x \in A$, there are arbitrarily small r such that $\frac{\lambda(B(x,r))}{\mu(B(x,r))} < a + \varepsilon$, $B(x,r) \subseteq V$. Therefore, by Corollary 9.12.3 again, there exists a disjoint sequence of these balls, $\{B_i\}$ satisfying this time,

$$\overline{\lambda}(A \setminus \cup_{i=1}^{\infty} B_i) = 0 \text{ and } \overline{\lambda}(A) \leq \sum_{i=1}^{\infty} \overline{\lambda}(A \cap B_i) \leq \sum_{i=1}^{\infty} \lambda(B_i)$$

$$\leq \sum_{i=1}^{\infty} (a+\varepsilon) \mu(B_i) \leq (a+\varepsilon) \mu(V) \leq (a+\varepsilon)(\varepsilon + \overline{\mu}(A))$$

Since $\varepsilon > 0$ is arbitrary, $\overline{\lambda}(A) \leq a\overline{\mu}(A)$.This proves the lemma in case A is bounded. If not, replace A with $A \cap B(\mathbf{0}, R)$, get the result for this and let $R \to \infty$. If $\overline{\mu}(A) = \infty$, there is nothing to show. ■

Theorem 11.5.3 *Let λ, μ be Radon measures on $\mathcal{F}_\lambda$ and $\mathcal{F}_\mu$ respectively. There exists a set of measure zero N containing Z such that for $x \notin N$, $D_\mu \lambda(x)$ exists and also $\mathcal{X}_{N^C}(\cdot) D_\mu \lambda(\cdot)$ is a μ measurable function. Furthermore, $D_\mu \lambda(x) < \infty$ μ a.e.*

Proof: First I show $D_\mu \lambda(x)$ exists a.e. Let $0 < a < b < \infty$ and let A be any bounded subset of

$$N(a,b) \equiv \{x \notin Z : \overline{D}_\mu \lambda(x) > b > a > \underline{D}_\mu \lambda(x)\}.$$

By Lemma 11.5.2, $a\overline{\mu}(A) \geq \overline{\lambda}(A) \geq b\overline{\mu}(A)$ and so $\overline{\mu}(A) = 0$ and so A is μ measurable. It follows $\mu(N(a,b)) = 0$ because $\mu(N(a,b)) \leq \sum_{m=1}^{\infty} \mu(N(a,b) \cap B(\mathbf{0}, m)) = 0$. Now

$$\{x \notin Z : \overline{D}_\mu \lambda(x) > \underline{D}_\mu \lambda(x)\} \subseteq \cup \{N(a,b) : 0 < a < b, \text{ and } a, b \in \mathbb{Q}\}$$

and the latter set is a countable union of sets of measure 0, so off a set of measure 0, N for which $N \supseteq Z$, one has $\overline{D}_\mu \lambda(x) = \underline{D}_\mu \lambda(x)$.

We can assume also that N is a Borel set from regularity considerations. See Theorem 9.11.2 for example. It remains to verify $\mathscr{X}_{N^C}(\cdot) D_\mu\lambda(\cdot)$ is finite a.e. and is μ measurable. Let $I = \{x : D_\mu\lambda(x) = \infty\}$. Then by Lemma 11.5.2

$$\overline{\lambda}(I \cap B(\mathbf{0},m)) \geq a\overline{\mu}(I \cap B(\mathbf{0},m))$$

for all $a > 0$, and since λ is finite on bounded sets, the above implies $\overline{\mu}(I \cap B(\mathbf{0},m)) = 0$ for each m which implies that I is μ measurable and has μ measure zero since $I = \cup_{m=1}^{\infty} I \cap B(\mathbf{0},m)$.

Now the issue is measurability. Let λ be an arbitrary Radon measure. I need show that $x \to \lambda(B(x,r))$ is measurable. Here is where it is convenient to have the balls be closed balls. If V is an open set containing $B(x,r)$, then for y close enough to x,$B(y,r) \subseteq V$ also and so, $\limsup_{y\to x} \lambda(B(y,r)) \leq \lambda(V)$. However, since V is arbitrary and λ is outer regular, or observing that $B(x,r)$ the closed ball is the intersection of nested open sets, it follows that $\limsup_{y\to x} \lambda(B(y,r)) \leq \lambda(B(x,r))$. Thus $x \to \lambda(B(x,r))$ is upper semicontinuous, similar for $x \to \mu(B(x,r))$ and so, $x \to \frac{\lambda(B(x,r))}{\mu(B(x,r))}$ is measurable. Hence $\mathscr{X}_{N^C}(x) D_\mu(\lambda)(x) = \lim_{r_i\to 0} \mathscr{X}_{N^C}(x) \frac{\lambda(B(x,r_i))}{\mu(B(x,r_i))}$ is also measurable. ■

Typically I will write $D_\mu\lambda(x)$ rather than the more precise $\mathscr{X}_{N^C}(x) D_\mu\lambda(x)$ since the values on the set of measure zero N are not important due to the completeness of the measure μ. This is done in the next section.

11.6 Radon Nikodym Theorem, Radon Measures

The Radon Nikodym theorem is an abstract result but this will be a special version. It will give a pointwise description in terms of the symmetric derivative of the Radon Nikodym derivative presented earlier.

Definition 11.6.1 *Let λ, μ be two Radon measures defined on $\mathscr{F}$, a σ algebra of subsets of an open set U. Then $\lambda \ll \mu$ means that whenever $\mu(E) = 0$, it follows that $\lambda(E) = 0$.*

Next is a representation theorem for λ in terms of an integral involving $D_\mu\lambda$.

Theorem 11.6.2 *Let λ and μ be Radon measures defined on $\mathscr{F}_\lambda, \mathscr{F}_\mu$ respectively, σ algebras of the open set U, then there exists a set of μ measure zero N such that $D_\mu\lambda(x)$ exists off N and if $E \subseteq N^C, E \in \mathscr{F}_\lambda \cap \mathscr{F}_\mu$, then $\lambda(E) = \int_U \left(D_\mu\lambda\right) \mathscr{X}_E d\mu$. If $\lambda \ll \mu$ on $\mathscr{F}_\lambda \cap \mathscr{F}_\mu$, then $\lambda(E) = \int_E D_\mu\lambda d\mu$. In any case, $\lambda(E) \geq \int_E D_\mu\lambda d\mu$ so $D_\mu\lambda$ is in $L^1_{loc}(\mathbb{R}^p, \mu)$ because $\lambda(B) < \infty$ for any ball B.*

Proof: The proof is based on Lemma 11.5.2. Let $E \subseteq N^C$ where N has μ measure 0 and includes the set Z along with the set where the symmetric derivative does not exist. It can be assumed that N is a G_δ set. Define

$$l_n(x) \equiv \sum_{k=1}^{\infty} a_{k-1}^n \mathscr{X}_{\left(D_\mu\lambda\right)^{-1}\left(I_k^n\right)}(x),\ u_n(x) \equiv \sum_{k=1}^{\infty} a_k^n \mathscr{X}_{\left(D_\mu\lambda\right)^{-1}\left(I_k^n\right)}(x)$$

where $I_k^n \equiv ((k-1)2^{-n}, k2^{-n}] \equiv (a_{k-1}^n, a_k^n]$ for $k,n \in \mathbb{N}$. Thus $u_n(x) \geq D_\mu\lambda(x) > l_n(x)$ and $u_n(x) - l_n(x) = 2^{-n}$. Also, $l_n(x)$ increases to $D_\mu\lambda(x)$. Letting

$$E_k^n \equiv \left[x \in E : D_\mu\lambda(x) \in I_k^n\right],$$

and assuming $\mu(E) < \infty$, $\int_E D_\mu \lambda d\mu \in [\int_E l_n d\mu, \int_E u_n d\mu]$

$$= \left[\sum_{k=1}^{\infty} a_{k-1}^n \mu\left(E_k^n\right), \sum_{k=1}^{\infty} a_k^n \mu\left(E_k^n\right)\right] \subseteq \left[\int_E l_n d\mu, \int_E l_n d\mu + 2^{-n}\mu(E)\right] \tag{11.7}$$

From Lemma 11.5.2, $\mu\left(E_k^n\right) a_k^n \geq \lambda\left(E_k^n\right) \geq a_{k-1}^n \mu\left(E_k^n\right)$ and so the interval in 11.7 contains $\sum_{k=1}^{\infty} \lambda\left(E_k^n\right)$. This equals $\lambda(E)$ because of Lemma 11.5.2 which implies

$$\begin{aligned} \lambda\left(E \cap \left\{x \in N^C : D_\mu \lambda(x) = 0\right\}\right) &\leq a\mu\left(E \cap \left\{x \in N^C : D_\mu \lambda(x) = 0\right\}\right) \\ &\leq a\mu(E),\ \mu(E) < \infty \end{aligned}$$

and since this is true for every positive a, it follows that

$$\lambda\left(E \cap \left\{x \in N^C : D_\mu \lambda(x) = 0\right\}\right) = 0$$

so the sum $\sum_{k=1}^{\infty} \lambda\left(E_k^n\right) = \lambda(E)$. Then, from the monotone convergence theorem in 11.7, one can pass to a limit and find that $\int_E D_\mu \lambda d\mu = \lambda(E)$.

Now if E is an arbitrary set in N^C, maybe not bounded, the above shows

$$\lambda(E \cap B(\mathbf{0}, n)) = \int_{E \cap B(\mathbf{0},n)} D_\mu \lambda d\mu$$

Let $n \to \infty$ and use the monotone convergence theorem. Thus for all $E \subseteq N^C$, $\lambda(E) = \int_E D_\mu \lambda d\mu$. For the last claim, $\int_E D_\mu \lambda d\mu = \int_{E \cap N^C} D_\mu \lambda d\mu = \lambda\left(E \cap N^C\right) \leq \lambda(E)$.

In case, $\lambda \ll \mu$, it does not matter that $E \subseteq N^C$ because, since $\mu(N) = 0$, so is $\lambda(N)$ and so

$$\lambda(E) = \lambda\left(E \cap N^C\right) = \int_{E \cap N^C} D_\mu \lambda d\mu = \int_E D_\mu \lambda d\mu$$

for any $E \in \mathscr{F}$. ■

What if λ and μ are just two arbitrary Radon measures defined on $\mathscr{F}$? What then? It was shown above that $D_\mu \lambda(x)$ exists for μ a.e. x, off a G_δ set N of μ measure 0 which includes Z, the set of x where $\mu(B(x,r)) = 0$ for some $r > 0$. Also, it was shown above that if $E \subseteq N^C$, then $\lambda(E) = \int_E D_\mu \lambda(x) d\mu$. Define for arbitrary $E \in \mathscr{F}$,

$$\lambda_\mu(E) \equiv \lambda\left(E \cap N^C\right),\ \lambda_\perp(E) \equiv \lambda(E \cap N)$$

Then

$$\begin{aligned} \lambda(E) &= \lambda(E \cap N) + \lambda\left(E \cap N^C\right) = \lambda_\perp(E) + \lambda_\mu(E) \\ &= \lambda(E \cap N) + \int_{E \cap N^C} D_\mu \lambda(x) d\mu \\ &= \lambda(E \cap N) + \int_E D_\mu \lambda(x) d\mu \equiv \lambda(E \cap N) + \lambda_\mu(E) \\ &\equiv \lambda_\perp(E) + \lambda_\mu(E) \end{aligned}$$

This shows the following corollary.

Corollary 11.6.3 *Let μ, λ be two Radon measures. Then there exist two measures, $\lambda_\mu, \lambda_\perp$ such that $\lambda_\mu \ll \mu$, $\lambda = \lambda_\mu + \lambda_\perp$ and a set of μ measure zero N such that $\lambda_\perp(E) = \lambda(E \cap N)$. Also λ_μ is given by the formula $\lambda_\mu(E) \equiv \int_E D_\mu \lambda(x) d\mu$.*

Proof: If $x \in N$, this could happen two ways, either $x \in Z$ or $D_\mu \lambda(x)$ fails to exist. It only remains to verify that λ_μ given above satisfies $\lambda_\mu \ll \mu$. However, this is obvious because if $\mu(E) = 0$, then $\int_E D_\mu \lambda(x)\, d\mu = 0$. ■

Since $D_\mu \lambda(x) = D_\mu \lambda(x)\, \mathcal{X}_{N^C}(x)$, it doesn't matter which we use but maybe $D_\mu \lambda(x)$ doesn't exist at some points of N, so although I will use $D_\mu \lambda(x)$, it might be more precise to use $D_\mu \lambda(x)\, \mathcal{X}_{N^C}(x)$.

This is sometimes called the Lebesgue decomposition.

How does this relate to Corollary 10.13.14? It tells how to find the function f in that Corollary as a symmetric derivative. This is very useful when you want to have an explicit description of the Radon Nikodym derivative.

11.7 Absolutely Continuous Functions

Can you integrate the derivative to get the function as in calculus? The answer is that sometimes you can and when this is the case, the function is called absolutely continuous. This is explained in this section. Recall the following which summarizes Theorems 9.10.1 on Page 263 and 9.7.4 on Page 256. In what follows m will be one dimensional Lebesgue measure. Recall that for F increasing, $F(x+) \equiv \lim_{h \to 0+} F(x+h)$, $F(x-) \equiv \lim_{h \to 0+} F(x-h)$.

Theorem 11.7.1 *Let F be an increasing function on $\mathbb{R}$. Then there is an outer measure μ and a σ algebra $\mathcal{F}$ on which μ is a measure such that $\mathcal{F}$ contains the Borel sets. This measure μ satisfies*

1. $\mu([a,b]) = F(b+) - F(a-)$,
2. $\mu((a,b)) = F(b-) - F(a+)$
3. $\mu((a,b]) = F(b+) - F(a+)$
4. $\mu([a,b)) = F(b-) - F(a-)$.

Furthermore, if E is any set in $\mathcal{F}$

$$\mu(E) = \sup\{\mu(K) : K \text{ compact, } K \subseteq E\} \tag{11.8}$$

$$\mu(E) = \inf\{\mu(V) : V \text{ is an open set } V \supseteq E\} \tag{11.9}$$

Of interest is the symmetric derivative of F, defined as

$$\lim_{h \to 0} \frac{F((x+h)+) - F((x-h)-)}{2h}$$

Now from Theorem 11.5.3 it follows that for a.e. x,

$$\lim_{h \to 0+} \frac{\mu([x-h, x+h])}{2h} = \lim_{h \to 0+} \frac{F((x+h)+) - F((x-h)-)}{2h} \equiv D_m \mu(x)$$

exists a.e. x. From Corollary 11.6.3, $\mu = \mu_\perp + \mu_m$ where $\mu_m(E) \equiv \int_E D_m \mu(x)\, dm$ and there is a Borel set of m measure zero N such that $\mu_\perp(E) = \mu(E \cap N)$

In case $\mu \ll m$ then from Theorem 11.6.2, if E is Borel, $\mu(E) = \int_E D_m \mu(x)\, dm$. This begs the following question.

What properties on F are equivalent to this measure μ being absolutely continuous with respect to m, $\mu \ll m$? Here is a definition of what it means for a **function** to be absolutely continuous with respect to Lebesgue measure. Thus there are now two things being called "absolutely continuous" functions and measures.

Definition 11.7.2 *Let $[a,b]$ be a closed and bounded interval and let $F:[a,b]\to\mathbb{R}$. Then F is said to be absolutely continuous if for every $\varepsilon>0$ there exists $\delta>0$ such that if $\sum_{i=1}^{m}|y_i-x_i|<\delta$ where the intervals (x_i,y_i) are non-overlapping, then it follows that $\sum_{i=1}^{m}|F(y_i)-F(x_i)|<\varepsilon$.*

It turns out that if, in the definition, you allow arbitrary intervals, non-overlapping or not, then you end up with a Lipschitz function which is obviously absolutely continuous. This is shown later.

The following theorem gives the desired equivalence between absolute continuity of F and $\mu\ll m$.

Theorem 11.7.3 *Let F be an increasing function on $[a,b]$ and let μ be the measure of Theorem 11.7.1. Then $\mu\ll m$ if and only if F is absolutely continuous.*

Proof: $\Rightarrow$First suppose that $\mu\ll m$. Then by Theorem 11.6.2, for all Borel sets E,

$$\mu(E)=\int_E D_m\mu(x)\,dm$$

In particular, F must be continuous because $D_m\mu$ is in L^1_{loc}. Thus

$$F(y-)-F(x+)=\int_{(x,y)}D_m\mu(x)\,dm=\int_{(x,y]}D_m\mu(x)\,dm=F(y+)-F(x+)$$

showing that for arbitrary y, $F(y-)=F(y+)$ so the function F is continuous as claimed. Also

$$F(b)-F(a)=\int_{[a,b]}D_m\mu(x)\,dm$$

so $D_m\mu$ is in $L^1([a,b],m)$.

If the function F is not absolutely continuous, then there exists $\varepsilon>0$ and open sets E_n consisting of unions of finitely many non-overlapping open intervals such that if $E_n=\cup_{i=1}^{m_n}(x_i^n,y_i^n)$, then $\sum_{i=1}^{m}|y_i^n-x_i^n|=m(E_n)<2^{-n}$ but

$$\int_{[a,b]}\mathcal{X}_{E_n}(x)\,D_m\mu(x)\,dm=\mu(E_n)=\sum_{i=1}^{m_n}\mu(x_i^n,y_i^n)=\sum_{i=1}^{m_n}(F(y_i^n)-F(x_i^n))\geq\varepsilon \tag{11.10}$$

However, $\mathcal{X}_{E_n}(x)\to 0$ a.e. because $\sum_n m(E_n)<\infty$ and so, by the Borel Cantelli lemma, there is a set of measure zero N such that for $x\notin N$, x is in only finitely many of the E_n. In particular, $\mathcal{X}_{E_n}(x)=0$ for all n large enough if $x\notin N$. Then by the dominated convergence theorem, the inequality 11.10 cannot be valid for all n because the limit of the integral on the left equals 0. This is a contradiction. Hence F must be absolutely continuous after all.

$\Leftarrow$Next suppose the function F is absolutely continuous. Suppose $m(E)=0$. Does it follow that $\mu(E)=0$? Let $\varepsilon>0$ be given. Let δ correspond to $\varepsilon/2$ in the definition of absolute continuity. Let $E\subseteq V$ where V is an open set such that $m(V)<\delta$. By Theorem 3.11.8, $V=\cup_i(a_i,b_i)$ where these open intervals are disjoint. It follows that for each n, $\frac{\varepsilon}{2}>\sum_{i=1}^{n}F(b_i)-F(a_i)=\mu\left(\cup_{i=1}^{n}(a_i,b_i)\right)$. Then letting $n\to\infty$, $\varepsilon>\frac{\varepsilon}{2}\geq\mu\left(\cup_{i=1}^{\infty}(a_i,b_i)\right)=\mu(V)\geq\mu(E)$. Since $\varepsilon>0$ is arbitrary, it follows that $\mu(E)=0$ and so $\mu\ll m$. ■

An example which shows that increasing and continuous is not enough, see Problem 5 on Page 273.

Corollary 11.7.4 *F is increasing on $[a,b]$ and absolutely continuous if and only if $F'(x)$ exists for a.e. x and F' is in $L^1([a,b],m)$ and for every x,y such that $a \le x \le y \le b$*

$$F(y) - F(x) = \int_x^y F'(t)\,dm$$

Proof: $\Rightarrow$Suppose first that F is absolutely continuous. Then by Theorem 11.6.2, for μ defined above, $\mu(E) = \int_E D_m\mu(x)\,dm$ for all E Borel. In particular,

$$F(y) - F(x) = \int_{(x,y)} D_m\mu(t)\,dm(t) \tag{11.11}$$

Since $D_m\mu$ is in $L^1([a,b],m)$, it follows that almost every point is a Lebesgue point and so for such Lebesgue points x,

$$\left|\frac{F(x+h)-F(x)}{h} - D_m\mu(x)\right| = \left|\frac{1}{h}\int_{[x,x+h]} (D_m\mu(t) - D_m\mu(x))\,dm(t)\right|$$

$$\le 2\left|\frac{1}{2h}\int_{[x-h,x+h]} |D_m\mu(t) - D_m\mu(x)|\,dm(t)\right|$$

which converges to 0 as $h \to 0$ since x is a Lebesgue point. Similarly, at each Lebesgue point, $\lim_{h\to 0}\frac{F(x)-F(x-h)}{h} = D_m\mu(x)$ Thus F is differentiable at each Lebesgue point and the derivative equals $D_m\mu$ at these points. Now 11.11 yields the desired result that the function can be recovered from integrating its derivative.

$\Leftarrow$Next suppose $F(y) - F(x) = \int_x^y F'(t)\,dm$ where $F'(t)$ exists a.e. and F' is in L^1. Then if $\{I_i\}_i$ are nonoverlapping intervals, $\int_{\cup_i I_i} F'(t)\,dm = m(\cup_i F(I_i)) < \varepsilon$ if $m(\cup_i I_i)$ is small enough, as an application of the dominated convergence theorem or as in the first part of Theorem 11.7.3. ■

The importance of the intervals being non overlapping is discussed in the following proposition. I think it is also interesting that it implies F is Lipschitz. In this proposition, F is defined on some interval, possibly $\mathbb{R}$.

Proposition 11.7.5 *Suppose Definition 11.7.2 is unchanged except do not insist that the intervals be non-overlapping. Then F is not just absolutely continuous, it is also Lipschitz. The converse also holds.*

Proof: Choose m such that $\sum_{n=m}^{\infty} 2^{-n} < \delta$ where δ goes with 1 in the definition of absolute continuity. Let $r \le 1/2$ and this implies that any such choice of r yields $\sum_{n=m}^{\infty} r^n < \delta$. Let $E_{nr} \equiv [F' > r^{-n}]$. If any E_{nr} has measure zero, then F' is bounded off a set of measure zero and this is what is desired. Otherwise, each E_{nr} has a point of density. Let $h_n \in [r^{n-1}, r^{n-2})$ so $r^{-n+1}h_n \ge 1$ and let r be small enough and m large enough that $\sum_{n=m}^{\infty} h_n < \delta$. For t_{N+m} a point of density for $E_{(N+m)r}$ let $I_n \equiv (t_{N+m} - rh_n/2, t_{N+m} + rh_n/2)$. Pick N large, say $N > 9$. Make r smaller if necessary so that for $n \in [m, N+m]$, $\frac{m(E_{nr}\cap I_n)}{m(I_n)} > \frac{1}{2}$. Note that $E_{(N+m)r} \subseteq E_{nr}$ for $n < N+m$ and $t_{N+m} \in E_{nr}$ must be a point of density for E_{nr}. Then, since $F' > r^{-n}$ on E_{nr}, one obtains the following sequence of inequalities.

$$4.5 < \frac{N+1}{2} = \sum_{n=m}^{m+N} \frac{1}{2} <$$

$$\sum_{n=m}^{m+N} \overbrace{r^{-n}}^{\geq 1} rh_n \frac{m(E_n \cap I_n)}{m(I_n)} = \sum_{n=m}^{m+N} r^{-n} m(I_n) \frac{m(E_n \cap I_n)}{m(I_n)} = \sum_{n=m}^{m+N} r^{-n} m(E_n \cap I_n)$$

$$\leq \sum_{n=m}^{m+N} \int_{t_{N+m}-rh_n/2}^{t_{N+m}+rh_n/2} \mathscr{X}_{E_n} F'(t)\, dm \leq \sum_{n=m}^{m+N} F(t_{N+m}+rh_n/2) - F(t_{N+m}-rh_n/2) < 1$$

by assumption, since the sum of the lengths of the intervals is smaller than δ. Thus $4.5 \leq 1$, a contradiction. Hence some E_{nr} has measure zero and so F' is bounded by a constant K off a set of measure zero. Hence, $|F(s)-F(t)| = \left|\int_s^t F'(u)\,du\right| \leq K|s-t|$ showing that F is Lipschitz.

The other direction is fairly obvious. If F is Lipschitz continuous, with Lipschitz constant K then if $\sum_{i=1}^m |x_{i+1}-x_i| < \delta$, then $\sum_{i=1}^m |F(x_{i+1})-F(x_i)| \leq K\sum_{i=1}^m |x_{i+1}-x_i|$ so if $\varepsilon > 0$ is given, let $\delta = \varepsilon/K$. ■

Note that when μ is the Lebesgue Stieltjes measure coming from increasing continuous F, it follows from the definition that $D_m\mu(x) = \lim_{h\to 0} \frac{F(x+h)-F(x-h)}{2h}$. The following is another characterization of absolute continuity.

Corollary 11.7.6 *Let μ be the Lebesgue Stieltjes measure described above for increasing F defined on $\mathscr{F}_\mu$ containing the Borel sets. Let $I \equiv \{x : D_m\mu(x) = \infty\}$. Then $\mu \ll m$ on $\mathscr{F}_m \cap \mathscr{F}_\mu$ if and only if $\mu(I)$ has measure 0. Here $\mathscr{F}_m$ is the σ algebra of Lebesgue measurable sets.*

Proof: $\Rightarrow$ If $\mu \ll m$, then by Theorem 11.6.2, for all $E \in \mathscr{F}_\mu \cap \mathscr{F}_m$,

$$\mu(E) = \int_E D_m\mu(x)\, dm$$

Then by the fundamental theorem of calculus, Theorem 11.6.2, there is a set of m measure zero N such that off this set, $D_m\mu(x)$ exists and is in $\mathbb{R}$. Thus $N \supseteq I$ and by absolute continuity, $\mu(I) = 0$.

$\Leftarrow$ Next suppose $\mu(I) = 0$. Then F has no jumps because if it did, then μ (a jump) > 0 and the jump is also contained in I. Let $m(E) = 0$ for E a bounded set. Then define

$$G_n \equiv \left\{ t : \liminf_{r\to 0} \frac{\mu(B(t,r))}{2r} \leq n \right\}, n \in \mathbb{N}$$

and note that $I = \left\{ t : \liminf_{r\to 0} \frac{\mu(B(t,r))}{2r} = \infty \right\}$. Consider $G_n \cap E$ where $m(E) = 0$. Let V be a bounded open set and $V \supseteq G_n \cap E$ and $m(V) < \varepsilon/n$. Then there is a Vitali cover of $G_n \cap E$ consisting of closed balls B such that $\mu(B) \leq nm(B)$, each ball contained in V. Then by the covering theorem Theorem 9.12.2 on Page 268, there is a disjoint union of these which covers $G_n \cap E$ called B_i, $\mu(E \cap G_n \setminus \cup_i B_i) = 0$ and so $\mu(E \cap G_n) = \sum_i \mu(B_i) \leq n\sum_i m(B_i) < nm(V) < \varepsilon$ and since ε is arbitrary, $\mu(E \cap G_n) = 0$. Now, since $\mu(I) = 0$,

$$\mu(E) = \mu(E \cap I) \cup \cup_n \mu(E \cap G_n) = \cup_n \mu(E \cap G_n) = \lim_{n\to\infty} \mu(E \cap G_n) = 0$$

In general, if $m(E) = 0$, let $E_n \equiv E \cap B(0,n)$. Then from what was just shown, $\mu(E_n) = 0$ and so, taking a limit, $\mu(E) = 0$ also. Thus $\mu \ll m$. ■

11.8 Total Variation

The total variation function of an absolutely continuous function is itself absolutely continuous. This is shown here along with some of its implications.

Definition 11.8.1 *A finite subset, P of $[a,b]$ is called a partition of $[x,y] \subseteq [a,b]$ if $P = \{x_0, x_1, \cdots, x_n\}$ where $x = x_0 < x_1 < \cdots, < x_n = y$. For $f : [a,b] \to \mathbb{R}$ and $P = \{x_0, x_1, \cdots, x_n\}$ define $V_P[x,y] \equiv \sum_{i=1}^{n} |f(x_i) - f(x_{i-1})|$. Denoting by $\mathscr{P}[x,y]$ the set of all partitions of $[x,y]$ define*

$$V[x,y] \equiv \sup_{P \in \mathscr{P}[x,y]} V_P[x,y].$$

For simplicity, $V[a,x]$ will be denoted by $V(x)$. It is called thetotal variation of the function f.

There are some simple facts about the total variation of an absolutely continuous function f which are contained in the next lemma.

Lemma 11.8.2 *Let f be an absolutely continuous function defined on $[a,b]$ and let V be its total variation function as described above. Then V is an increasing bounded function. Also if P and Q are two partitions of $[x,y]$ with $P \subseteq Q$, then $V_P[x,y] \leq V_Q[x,y]$ and if $[x,y] \subseteq [z,w]$,*

$$V[x,y] \leq V[z,w] \tag{11.12}$$

If $P = \{x_0, x_1, \cdots, x_n\}$ is a partition of $[x,y]$, then

$$V[x,y] = \sum_{i=1}^{n} V[x_i, x_{i-1}]. \tag{11.13}$$

Also if $y > x$,

$$V(y) - V(x) \geq |f(y) - f(x)| \tag{11.14}$$

and the function, $x \to V(x) - f(x)$ is increasing. The total variation function V is absolutely continuous.

Proof: The claim that V is increasing is obvious as is the next claim about $P \subseteq Q$ leading to $V_P[x,y] \leq V_Q[x,y]$. To verify this, simply add in one point at a time and verify that from the triangle inequality, the sum involved gets no smaller. The claim that V is increasing consistent with set inclusion of intervals is also clearly true and follows directly from the definition.

Now let $t < V[x,y]$ where $P_0 = \{x_0, x_1, \cdots, x_n\}$ is a partition of $[x,y]$. There exists a partition, P of $[x,y]$ such that $t < V_P[x,y]$. Without loss of generality it can be assumed that $\{x_0, x_1, \cdots, x_n\} \subseteq P$ since if not, you can simply add in the points of P_0 and the resulting sum for the total variation will get no smaller. Let P_i be those points of P which are contained in $[x_{i-1}, x_i]$. Then $t < V_p[x,y] = \sum_{i=1}^{n} V_{P_i}[x_{i-1}, x_i] \leq \sum_{i=1}^{n} V[x_{i-1}, x_i]$. Since $t < V[x,y]$ is arbitrary,

$$V[x,y] \leq \sum_{i=1}^{n} V[x_i, x_{i-1}] \tag{11.15}$$

Note that 11.15 does not depend on f being absolutely continuous.

Suppose now that f is absolutely continuous. Let δ correspond to $\varepsilon = 1$. Then if $[x,y]$ is an interval of length no larger than δ, the definition of absolute continuity implies $V[x,y] < 1$. Then from 11.15

$$V[a,n\delta] \leq \sum_{i=1}^{n} V[a+(i-1)\delta, a+i\delta] < \sum_{i=1}^{n} 1 = n.$$

Thus V is bounded on $[a,b]$. Now let P_i be a partition of $[x_{i-1},x_i]$ such that $V_{P_i}[x_{i-1},x_i] > V[x_{i-1},x_i] - \frac{\varepsilon}{n}$. Then letting $P = \cup P_i$,

$$-\varepsilon + \sum_{i=1}^{n} V[x_{i-1},x_i] < \sum_{i=1}^{n} V_{P_i}[x_{i-1},x_i] = V_P[x,y] \leq V[x,y].$$

Since ε is arbitrary, 11.13 follows from this and 11.15.

Now let $x < y$. $V(y) - f(y) - (V(x) - f(x)) =$

$$V(y) - V(x) - (f(y) - f(x)) \geq V(y) - V(x) - |f(y) - f(x)| \geq 0.$$

It only remains to verify that V is absolutely continuous.

Let $\varepsilon > 0$ be given and let δ correspond to $\varepsilon/2$ in the definition of absolute continuity applied to f. Suppose $\sum_{i=1}^{n} |y_i - x_i| < \delta$ and consider $\sum_{i=1}^{n} |V(y_i) - V(x_i)|$. By 11.15 this last is no larger than $\sum_{i=1}^{n} V[x_i,y_i]$. Now let P_i be a partition of $[x_i,y_i]$ such that $V_{P_i}[x_i,y_i] + \frac{\varepsilon}{2n} > V[x_i,y_i]$. Then by the definition of absolute continuity,

$$\sum_{i=1}^{n} |V(y_i) - V(x_i)| = \sum_{i=1}^{n} V[x_i,y_i] \leq \sum_{i=1}^{n} V_{P_i}[x_i,y_i] + \eta < \varepsilon/2 + \varepsilon/2 = \varepsilon$$

and shows V is absolutely continuous as claimed. ■

Now with the above results, the following is the main result on absolutely continuous functions.

Theorem 11.8.3 *Let $f : [a,b] \to \mathbb{R}$ be a function. Then f is absolutely continuous if and only if $f'(t)$ exists a.e., f' is in $L^1([a,b],m)$, and for every $a \leq x \leq y \leq b$,*

$$f(y) - f(x) = \int_x^y f'(t)\,dt \equiv \int_{[x,y]} f'(t)\,dm(t)$$

Proof: Suppose f is absolutely continuous. Using Lemma 11.8.2, $f(x) = V(x) - (V(x) - f(x))$, the difference of two increasing functions, both of which are absolutely continuous. See Problem 2 on Page 334. Denote the derivatives of these two increasing functions by k and l respectively. Then for $x \leq y$,

$$f(y) - f(x) = \int_{[x,y]} k(t)\,dm(t) - \int_{[x,y]} l(t)\,dm(t)$$

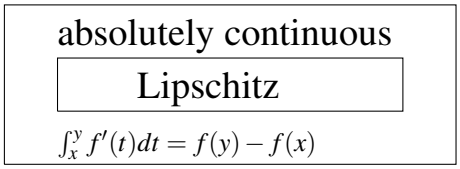
Letting $g(t) \equiv k(t) - l(t)$, it follows that $f(y) - f(x) = \int_x^y g(t)\,dt$ where $g \in L^1$. Then from the fundamental theorem of calculus, Theorem 11.3.2, if x is a Lebesgue point of g, not equal to one of the end points.

$$\left| \frac{f(x+h) - f(x)}{h} - g(x) \right| = \left| \frac{1}{h} \int_x^{x+h} g(t) - g(x)\,dt \right| \leq 2\left(\frac{1}{2h} \int_{x-h}^{x+h} |g(t) - g(x)|\,dt \right)$$

which converges to 0. Hence $g(x) = f'(x)$ a.e.

Next suppose $f(y) - f(x) = \int_x^y f'(t)\,dt$ where $f' \in L^1$. If f is not absolutely continuous, there exists $\varepsilon > 0$ and sets V_n each of which is the union of non-overlapping intervals such that $m(V_n) < 2^{-n}$ but $\int_{V_n} |f'(t)|\,dt \geq \varepsilon$. However, by the Borel Cantelli lemma, there exists a set of measure zero N such that for $x \notin N$, it follows that x is in only finitely many of the V_n. Thus $\mathcal{X}_{V_n}(x) \to 0$. Then a use of the dominated convergence theorem implies that $\lim_{n\to\infty} \int_{V_n} |f'(t)|\,dt = 0$ which is a contradiction. Thus f must be absolutely continuous. ■

The following picture illustrates the main items shown so far about functions of one variable.



11.9 Constructing Measures from Functionals

Here is a theorem which is the main result on measures and functionals defined on a space of continuous functions. The typical situation is of a metric space in which closed balls are compact.

Definition 11.9.1 *$C_c(X)$ will denote the complex values functions which have compact support in some metric space X. This is clearly a linear space. Then a linear function $L : C_c(X) \to \mathbb{C}$ is called "postitive" if whenever $f \geq 0$, then $Lf \geq 0$.*

Theorem 11.9.2 *Let $L : C_c(X) \to \mathbb{C}$ be a positive linear functional where X is a metric space and X is a countable union of compact sets. Then there exists a complete measure μ defined on a σ algebra $\mathcal{F}$ which contains the Borel sets $\mathcal{B}(X)$ which is finite on compact sets and has the following properties.*

1. *μ is regular and if E is measurable, there are F_σ and G_δ sets F, G such that $F \subseteq E \subseteq G$ and $\mu(G \setminus F) = 0$.*
2. *For all $f \in C_c(X)$, $Lf = \int_X f\,d\mu$*

Proof: See the notation and lemmas near Definition 3.12.3 having to do with partitions of unity on a metric space for what is needed in this proof. For V open, let $\bar{\mu}(V) \equiv \sup\{Lf : f \prec V\}$. Then for an arbitrary set F, let $\bar{\mu}(F) \equiv \inf\{\bar{\mu}(V) : V \supseteq F\}$, $\bar{\mu}(\emptyset) \equiv 0$. In what follows, V will be an open set and K a compact set.

Claim 1: $\bar{\mu}$ is well defined.

Proof of Claim 1: Note there are two descriptions of $\bar{\mu}(V)$ for V open. They need to be the same. Let $\bar{\mu}_1$ be the definition involving supremums of Lf and let $\bar{\mu}$ be the general definition. Let $V \subseteq U$ where V, U open. Then by definition, $\bar{\mu}(V) \leq \bar{\mu}_1(U)$ and so $\bar{\mu}(V) \equiv \inf\{\bar{\mu}_1(U) : U \supseteq V\} \geq \bar{\mu}_1(V)$. However, $V \subseteq V$ and so $\bar{\mu}(V) \leq \bar{\mu}_1(V)$. ■

Claim 2: $\bar{\mu}$ is finite on compact sets. Also, if $K \prec f$, it follows that $\bar{\mu}(K) \leq L(f) < \infty$.

Proof of Claim 2: Let $K \prec f \prec X$. Let $V_\varepsilon \equiv \{x : f(x) > 1 - \varepsilon\}$, an open set since f is continuous. Then let $g \prec V_\varepsilon$ so it follows that $\frac{f}{1-\varepsilon} \geq g$. Then $L(g) \leq \frac{1}{1-\varepsilon} L(f) < \infty$. Then

taking the sup over all such g, it follows that $\bar{\mu}(K) \leq \bar{\mu}(V_\varepsilon) \leq \frac{1}{1-\varepsilon} Lf$. Now let $\varepsilon \to 0$ and conclude that $\bar{\mu}(K) \leq L(f)$. ■

Claim 3: $\bar{\mu}$ is subadditive: $\bar{\mu}(\cup_i E_i) \leq \sum_i \bar{\mu}(E_i)$.

Proof of Claim 3: First consider the case of open sets. Let $V = \cup_i V_i$. Let $l < \bar{\mu}(V)$. Then there exists $f \prec V$ with $Lf > l$. Then $\sup(f)$ is contained in $\cup_{i=1}^n V_i$. Now let $\sup \psi_i \subseteq V_i$ and $\sum_{i=1}^n \psi_i = 1$ on $\sup(f)$. This is from Theorem 3.12.5. Then

$$l < Lf = \sum_{i=1}^n L(\psi_i f) \leq \sum_{i=1}^n \bar{\mu}(V_i) \leq \sum_i \bar{\mu}(V_i).$$

Since l is arbitrary, it follows that $\bar{\mu}(V) \leq \sum_i \bar{\mu}(V_i)$. Now consider the general case. Let $E = \cup_i E_i$. If $\sum_i \bar{\mu}(E_i) = \infty$, there is nothing to show. Assume then that this sum is finite and let $V_i \supseteq E_i, \bar{\mu}(E_i) + \frac{\varepsilon}{2^i} > \bar{\mu}(V)$. Then

$$\bar{\mu}(E) \leq \bar{\mu}(\cup_i V_i) \leq \sum_i \bar{\mu}(V_i) \leq \sum_i \left(\bar{\mu}(E_i) + \frac{\varepsilon}{2^i}\right) = \sum_i \bar{\mu}(E_i) + \varepsilon$$

Since ε is arbitrary, this shows $\bar{\mu}$ is subadditive. ■

Claim 4: If $\operatorname{dist}(A,B) = \delta > 0$, then $\bar{\mu}(A \cup B) = \bar{\mu}(A) + \bar{\mu}(B)$.

Proof of Claim 4: If the right side is infinite, there is nothing to show so we can assume that $\bar{\mu}(A), \bar{\mu}(B)$ are both finite. First suppose U, V are open and disjoint having finite outer measure. Let $\bar{\mu}(U) \leq Lf_1 + \varepsilon$ where $f_1 \prec U$ and let $f_2 \prec V$ with $\bar{\mu}(V) \leq L(f_2) + \varepsilon$. Then

$$\bar{\mu}(U \cup V) \leq \bar{\mu}(U) + \bar{\mu}(V) \leq Lf_1 + Lf_2 + 2\varepsilon \leq L(f_1 + f_2) + 2\varepsilon \leq \bar{\mu}(U \cup V) + 2\varepsilon$$

Since ε is arbitrary, this shows that $\bar{\mu}(U \cup V) = \bar{\mu}(U) + \bar{\mu}(V)$. Now in case A, B are as assumed, let $U \equiv \cup_{x \in U} B(x, \delta/3), V \equiv \cup_{x \in V} B(x, \delta/3)$. Then these are disjoint open sets containing A and B respectively. Then there is O open, $O \supseteq A \cup B$ such that $\bar{\mu}(A \cup B) + \varepsilon > \bar{\mu}(O)$. Replacing U with $U \subseteq O$ and V with $V \cap O$, we can assume $\bar{\mu}(A \cup B) + \varepsilon > \bar{\mu}(U \cup V)$. Then

$$\begin{aligned} \bar{\mu}(A) + \bar{\mu}(B) &\leq \bar{\mu}(U) + \bar{\mu}(V) = \bar{\mu}(U \cup V) \\ &< \varepsilon + \bar{\mu}(A \cup B) \leq \varepsilon + \bar{\mu}(A) + \bar{\mu}(B) \end{aligned}$$

Since ε is arbitrary, this shows that $\bar{\mu}(A) + \bar{\mu}(B) = \bar{\mu}(A \cup B)$.

From Theorem 9.8.4 there is a complete measure μ defined on a σ algebra $\mathscr{F}$ which equals $\bar{\mu}$ on $\mathscr{F}$. From Claim 4 and Theorem 9.9.1, $\mathscr{F}$ contains the Borel sets $\mathscr{B}(X)$. From the definition, μ is outer regular and so it follows from Theorem 9.4.6 that μ is regular because it is finite on compact sets and X is the union of countably many compact sets so μ is σ finite. Hence, by that theorem, the claimed approximation result also holds.

It only remains to show that the integrals with respect to the measure represent the functional. This will complete the proof.

Claim 5: $\int f d\mu = Lf$ for all $f \in C_c(X)$.

Proof: Let $f \in C_c(X)$, f real-valued, and suppose $f(X) \subseteq [a,b]$. Choose $t_0 < a$ and let $t_0 < t_1 < \cdots < t_n = b$, $t_i - t_{i-1} < \varepsilon$. Let

$$E_i = f^{-1}((t_{i-1}, t_i]) \cap \operatorname{spt}(f). \tag{11.16}$$

Note that $\cup_{i=1}^n E_i$ is a closed set equal to $\operatorname{spt}(f)$.

$$\cup_{i=1}^n E_i = \operatorname{spt}(f) \tag{11.17}$$

Since $X = \cup_{i=1}^n f^{-1}((t_{i-1}, t_i])$. Let $V_i \supseteq E_i, V_i$ is open and let V_i satisfy

$$f(x) < t_i + \varepsilon \text{ for all } x \in V_i,\ \mu(V_i \setminus E_i) < \varepsilon/n. \tag{11.18}$$

By Theorem 3.12.5, there exists $h_i \in C_c(X)$ such that

$$h_i \prec V_i,\ \sum_{i=1}^n h_i(x) = 1 \text{ on } \operatorname{spt}(f).$$

Now note that for each i, $f(x)h_i(x) \le h_i(x)(t_i+\varepsilon)$. (If $x \in V_i$, this follows from 19.11. If $x \notin V_i$ both sides equal 0.) Therefore,

$$\begin{aligned} Lf &= L(\sum_{i=1}^n f h_i) \le L(\sum_{i=1}^n h_i(t_i+\varepsilon)) = \sum_{i=1}^n (t_i+\varepsilon) L(h_i) \\ &= \sum_{i=1}^n (|t_0| + t_i + \varepsilon) L(h_i) - |t_0| L\left(\sum_{i=1}^n h_i\right). \end{aligned}$$

Now note that $|t_0| + t_i + \varepsilon \ge 0$ and so from the definition of μ and Lemma 19.7.5, this is no larger than

$$\begin{aligned} &\sum_{i=1}^n (|t_0| + t_i + \varepsilon)\mu(V_i) - |t_0|\mu(\operatorname{spt}(f)) \\ &\le \sum_{i=1}^n (|t_0| + t_i + \varepsilon)(\mu(E_i) + \varepsilon/n) - |t_0|\mu(\operatorname{spt}(f)) \\ &\le |t_0| \overbrace{\sum_{i=1}^n \mu(E_i)}^{\mu(\operatorname{spt}(f))} + \frac{\varepsilon}{n} n |t_0| + \sum_i t_i \mu(E_i) \\ &\quad + \sum_i t_i \frac{\varepsilon}{n} + \sum_i \varepsilon \mu(E_i) + \frac{\varepsilon^2}{n} - |t_0| \mu(\operatorname{spt}(f)) \\ &\le \varepsilon |t_0| + \varepsilon(|t_0| + |b|) + \varepsilon\mu(\operatorname{spt}(f)) + \varepsilon^2 + \sum_i t_i \mu(E_i) \\ &\le \varepsilon |t_0| + \varepsilon(|t_0| + |b|) + 2\varepsilon\mu(\operatorname{spt}(f)) + \varepsilon^2 + \sum_{i=1}^n t_{i-1}\mu(E_i) \\ &\le \varepsilon(2|t_0| + |b| + 2\mu(\operatorname{spt}(f)) + \varepsilon) + \int f d\mu \end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, $Lf \le \int f d\mu$ for all $f \in C_c(X)$, f real. Hence equality holds because $L(-f) \le -\int f d\mu$ so $L(f) \ge \int f d\mu$. Thus $Lf = \int f d\mu$ for all $f \in C_c(X)$. Just apply the result for real functions to the real and imaginary parts of f. ■

Using Corollary 9.4.9 we obtain the following corollary. Note that the conditions of the above theorem imply that X is a Polish space in the usual case where closed balls are compact.

Corollary 11.9.3 *If X is a Polish space then in the above theorem, we obtain inner regularity of μ in terms of compact sets. That is if $F \in \mathcal{F}$, then*

$$\mu(F) = \sup\{\mu(K) : K \subseteq F, K \text{ compact}\}$$

11.10 The p Dimensional Lebesgue Measure

Theorem 11.9.2 will provide many examples of Radon measures on $\mathbb{R}^p$. Lebesgue measure is obtained by letting

$$Lf \equiv \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f(x_1,...,x_p)\, dm_1(x_1) \cdots dm_p(x_p)$$

for $f \in C_c(\mathbb{R}^p)$. Thus Lebesgue measure is a Radon measure, denoted as m_p. In this case, the σ algebra will be denoted as $\mathscr{F}_p$. Lebesgue measure also has other very important properties. Integrals can be easily computed and the measure is translation invariant.

Theorem 11.10.1 *Whenever f is measurable and nonnegative, then whenever g is Borel measurable and equals f a.e. and h is Borel and equals f a.e.*

$$\int_{\mathbb{R}} \cdots \int_{\mathbb{R}} h(x_1,...,x_p)\, dm_1(x_{i_1}) \cdots dm_p(x_{i_p}) =$$

$$\int_{\mathbb{R}^p} f dm_p = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} g(x_1,...,x_p)\, dm_1(x_{i_1}) \cdots dm_p(x_{i_p})$$

where $(i_1, i_2, ..., i_p)$ is any permutation of the integers $\{1, 2, ..., p\}$. Also, m_p is regular and complete. If R is of the form $\prod_{i=1}^p I_i$ where I_i is an interval, then $m_p(R)$ is the product of the lengths of the sides of R. Also if $E \in \mathscr{F}_p$, then $m_p(\boldsymbol{x} + E) = m_p(E)$.

Proof: Let $\mathscr{K}$ consist of all open rectangles $\prod_i (a_i, b_i)$ along with $\emptyset$ and $\mathbb{R}^p$. Thus this is a π system. Let $R_n \equiv \prod_{i=1}^p (-n, n)$. Let $\mathscr{G}$ consist of the Borel sets $E \subseteq \mathbb{R}^p$ such that $\int_{R_n \cap E} dm_p = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \mathscr{X}_{R_n \cap E} dm_1(x_{i_1}) \cdots dm_p(x_{i_p})$. Then $\mathscr{K} \subseteq \mathscr{G}$. It is obvious from the monotone convergence theorem that $\mathscr{G}$ is closed with respect to countable disjoint unions. Indeed, this theorem implies that for $E = \cup_i E_i$, the E_i disjoint,

$$\begin{aligned}
\int_{R_n \cap E} dm_p &= \lim_{m\to\infty} \int_{R_n \cap \cup_{i=1}^m E_i} dm_p = \lim_{m\to\infty} \left(\sum_{i=1}^m \int \mathscr{X}_{R_n \cap E_i} dm_p \right) \\
&= \lim_{m\to\infty} \left(\sum_{i=1}^m \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \mathscr{X}_{R_n \cap E_i} dm_1(x_{i_1}) \cdots dm_p(x_{i_p}) \right) \\
&= \lim_{m\to\infty} \left(\int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \sum_{i=1}^m \mathscr{X}_{R_n \cap E_i} dm_1(x_{i_1}) \cdots dm_p(x_{i_p}) \right) \\
&= \lim_{m\to\infty} \left(\int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \mathscr{X}_{R_n \cap \cup_{i=1}^m E_i} dm_1(x_{i_1}) \cdots dm_p(x_{i_p}) \right) \\
&= \left(\int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \mathscr{X}_{R_n \cap E} dm_1(x_{i_1}) \cdots dm_p(x_{i_p}) \right)
\end{aligned}$$

As to complements, $\int_{R_n} dm_p = \int_{R_n \cap E^C} dm_p + \int_{R_n \cap E} dm_p$. Thus

$$\int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \mathscr{X}_{R_n \cap E^C} dm_1(x_{i_1}) \cdots dm_p(x_{i_p}) =$$

$$\int_{\mathbb{R}} \cdots \int_{\mathbb{R}} (\mathscr{X}_{R_n} - \mathscr{X}_{R_n \cap E}) dm_1(x_{i_1}) \cdots dm_p(x_{i_p}) = \int_{R_n \cap E^C} dm_p$$

It follows that $\mathscr{G} = \mathscr{B}(\mathbb{R}^p)$, the Borel sets. Hence

$$\int_{\mathbb{R}^p} \mathscr{X}_E dm_p = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \mathscr{X}_E dm_1(x_{i_1}) \cdots dm_p(x_{i_p})$$

for any Borel set E after letting $n \to \infty$ and using the monotone convergence theorem. Approximating a nonnegative Borel function g with an increasing sequence of simple Borel measurable functions, and using the monotone convergence theorem yields

$$\int g dm_p = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} g(x_1, ..., x_p) dm_1(x_{i_1}) \cdots dm_p(x_{i_p})$$

The claim about the measure of a box being the product of the lengths of its sides also comes from this.

By Proposition 11.1.2, for f measurable, there exists g Borel measurable such that $g = f$ a.e. and $g \leq f$. Then

$$\int_{\mathbb{R}^p} f dm_p = \int_{\mathbb{R}^p} g dm_p = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} g(x_1, ..., x_p) dm_1(x_{i_1}) \cdots dm_p(x_{i_p})$$

It is similar if $h \geq f$ and equal to f a.e.

It remains to consider the claim about translation invariance. If R is a box, $R = \prod_{i=1}^p (a_i, b_i)$, then it is clear that $m_p(\boldsymbol{x} + R) = m_p(R)$. Let $\mathscr{K}$ be as above and let $\mathscr{G}$ be those Borel sets E for which $m_p(\boldsymbol{x} + E \cap R_n) = m_p(E \cap R_n)$ where R_n is as above. Thus $\mathscr{G}$ contains $\mathscr{K}$. Then it is obvious $\mathscr{G}$ is closed with respect to countable disjoint unions. The case of complements maybe is not as obvious.

$$(\boldsymbol{x} + R_n) \setminus (\boldsymbol{x} + R_n \cap E) = \boldsymbol{x} + R_n \cap E^C.$$

Then

$$\begin{aligned} m_p(\boldsymbol{x} + R_n \cap E^C) &= m_p(\boldsymbol{x} + R_n) - m_p(\boldsymbol{x} + R_n \cap E) \\ &= m_p(R_n) - m_p(R_n \cap E) = m_p(R_n \cap E^C) \end{aligned}$$

Thus by Dynkin's lemma, $\mathscr{G} = \mathscr{B}(\mathbb{R}^p)$. Thus for all E Borel,

$$m_p(E \cap R_n) = m_p(\boldsymbol{x} + E \cap R_n).$$

Now let $n \to \infty$. It follows that m_p is translation invariant for all Borel sets.

In general, if E is Lebesgue measurable, it follows from Proposition 11.1.2 that there are sets $F \subseteq E \subseteq G$ where F, G are Borel and $m_p(F) = m_p(E) = m_p(G)$. Then

$$m_p(E) = m_p(F) = m_p(F + \boldsymbol{x}) \leq m_p(E + \boldsymbol{x}) \leq m_p(G + \boldsymbol{x}) = m_p(G) = m_p(E)$$

and so all the inequalities are equal signs. Hence $m_p(E + \boldsymbol{x}) = m_p(E)$. ■

The following is a useful lemma. In this lemma, X_i will be some metric space or more generally a topological space. It is useful in recognizing a Borel measurable set when you see it.

Lemma 11.10.2 *If E_i is a Borel set in X_i, then $\prod_{k=1}^p E_{i_k}$ is a Borel set in $\prod_{k=1}^p X_{i_k}$.*

Proof: Let $\pi_{i_r} : \prod_{k=1}^{p} X_{i_k} \to X_{i_r}$ be the projection map. That is $\pi_{i_r}(\boldsymbol{x}) = x_{i_r}$ when $\boldsymbol{x} = \left(x_{i_1}, x_{i_2}, ..., x_{i_p}\right)$. Obviously this is continuous. Therefore, if U is an open set in $X_{i_r}, \pi_{i_r}^{-1}(U) = X_{i_1} \times X_{i_2} \times \cdots \times U \times \cdots \times X_{i_p}$. Is an open set. Let $\mathscr{B}_{i_r}$ be the Borel sets of X_{i_r} E such that $\pi_{i_r}^{-1}(E) = X_{i_1} \times X_{i_2} \times \cdots \times E \times \cdots \times X_{i_p}$ is a Borel set in $\prod_{k=1}^{p} X_{i_k}$. Then $\mathscr{B}_{i_r}$ is a σ algebra and it contains the open sets. Therefore, it contains the Borel sets of X_{i_r}. It follows that $\prod_{k=1}^{p} E_{i_k} = \cap_{k=1}^{p} \pi_{i_k}^{-1}\left(E_{i_k}\right)$ is a finite intersection of Borel sets in $\prod_{k=1}^{p} X_{i_k}$ and so it is also a Borel set. ■

Example 11.10.3 *Let $A \equiv \{(x,y) : y < g(x)\}$ for g a Borel measurable real valued function. Then A is a Borel set.*

To see this, partition $\mathbb{R}$ into equally spaced points $\left\{r_k^n\right\}_{k=-\infty}^{\infty}, r_k^n < r_{k+1}^n, r_{k+1}^n - r_k^n = 2^{-n}$ and let $g_n(x) \equiv \sum_{k=-\infty}^{\infty} r_{k-1}^n \mathscr{X}_{g^{-1}\left((r_{k-1}^n, r_k^n]\right)}(x)$ so that $g_n(x) \to g(x)$ for each x. Let $A_n \equiv \{(x,y) : y < g_n(x)\}$. Now each A_k is Borel by the above Lemma. Then thanks to convergence, $A = \cap_{m=1}^{\infty} \cap_{k \geq m} A_k$ so A is Borel.

Example 11.10.4 *Let $A \equiv \{(x,y) : y \leq g(x)\}$ for g a Borel measurable function. Then A is a Borel set.*

This follows from observing that if $A_n \equiv \{(x,y) : y < g(x) + 2^{-n}\}$ then $A = \cap_{n=1}^{\infty} A_n$. Thus sets of the form $[a,b] \times \{(x,y) : y \leq g(x)\}$ for g Borel measurable are Borel measurable. These examples justify the usual calculus manipulations involving iterated integrals, the next example being an illustration of this.

Example 11.10.5 *Find the iterated integral $\int_0^1 \int_x^1 \frac{\sin(y)}{y} dy dx$.*

Notice the limits. The iterated integral equals $\int_{\mathbb{R}^2} \mathscr{X}_A(x,y) \frac{\sin(y)}{y} dm_2$ where

$$A = \{(x,y) : x \leq y \text{ where } x \in [0,1]\}.$$

Fubini's theorem can be applied because the function $(x,y) \to \sin(y)/y$ is continuous except at $y = 0$ and can be redefined to be continuous there. The function is also bounded so $(x,y) \to \mathscr{X}_A(x,y) \frac{\sin(y)}{y}$ clearly is in $L^1\left(\mathbb{R}^2\right)$. Therefore,

$$\int_{\mathbb{R}^2} \mathscr{X}_A(x,y) \frac{\sin(y)}{y} dm_2 = \int \int \mathscr{X}_A(x,y) \frac{\sin(y)}{y} dx dy$$

$$= \int_0^1 \int_0^y \frac{\sin(y)}{y} dx dy = \int_0^1 \sin(y)\, dy = 1 - \cos(1)$$

Here is a general and important result.

Example 11.10.6 *Integration by parts. Suppose f, g are both absolutely continuous, then*

$\int_a^b fg'dt =$

$$\begin{aligned}
&\int_a^b \left(f(a) + \int_a^t f'(s)\,ds \right) g'(t)\,dt \\
= \;& f(a)(g(b)-g(a)) + \int_a^b \int_a^t f'(s)\,ds g'(t)\,dt \\
= \;& f(a)(g(b)-g(a)) + \int_a^b \int_s^b f'(s)\,g'(t)\,dtds \\
= \;& f(a)(g(b)-g(a)) + \int_a^b f'(s)(g(b)-g(s))\,ds \\
= \;& f(a)(g(b)-g(a)) + g(b)(f(b)-f(a)) - \int_a^b f'(s)\,g(s)\,ds \\
= \;& g(b)f(b) - f(a)g(a) - \int_a^b f'(s)\,g(s)\,ds
\end{aligned}$$

Next is an easy corollary for the fundamental theorem of calculus in the special case where $\mu = m_p$. This kind of thing will not hold for μ a general Radon measure.

Corollary 11.10.7 *Let $f \in L^1(\mathbb{R}^p, m_p)$ meaning that ($\int_\Omega |f|\,dm_p < \infty$). Then for $\boldsymbol{x} \in \mathbb{R}^p$, let $Q(\boldsymbol{x},r) \equiv \prod_{j=1}^p I_j$ where I_j is an interval, open or closed, having x_j at one end point. Then for a.e. $\boldsymbol{x}$,*

$$\lim_{r\to 0} \frac{1}{m_p(Q(\boldsymbol{x},r))} \int_{Q(\boldsymbol{x},r)} |f(\boldsymbol{y}) - f(\boldsymbol{x})|\,dm_p(y) = 0$$

In particular, $\lim_{r\to 0} \frac{1}{m_p(Q(\boldsymbol{x},r))} \int_{Q(\boldsymbol{x},r)} f(\boldsymbol{y})\,dm_p(y) = f(\boldsymbol{x})$ *for a.e. $\boldsymbol{x}$.*

Proof: Let $B(\boldsymbol{x},r)$ be the ball defined using $\|\cdot\|_\infty$ having $\boldsymbol{x}$ at the center. Thus this ball consists of 2^p translated copies of $Q(\boldsymbol{x},r)$ and so

$$\begin{aligned}
&\frac{1}{m_p(Q(\boldsymbol{x},r))} \int_{Q(\boldsymbol{x},r)} |f(\boldsymbol{y}) - f(\boldsymbol{x})|\,dm_p(y) \\
\leq \;& 2^p \frac{1}{m_p(B(\boldsymbol{x},r))} \int_{B(\boldsymbol{x},r)} |f(\boldsymbol{y}) - f(\boldsymbol{x})|\,dm_p(y)
\end{aligned}$$

and the latter converges to 0 as $r \to 0$. ■

11.11 The Brouwer Fixed Point Theorem

I found this proof of the Brouwer fixed point theorem in Evans [18] and Dunford and Schwartz [16]. The main idea which makes proofs like this work is Lemma 7.11.2 which is stated next for convenience.

Lemma 11.11.1 *Let $\boldsymbol{g} : U \to \mathbb{R}^p$ be C^2 where U is an open subset of $\mathbb{R}^p$. Then*

$$\sum_{j=1}^p \operatorname{cof}(D\boldsymbol{g})_{ij,j} = 0,$$

where here $(D\boldsymbol{g})_{ij} \equiv g_{i,j} \equiv \frac{\partial g_i}{\partial x_j}$. Also, $\operatorname{cof}(D\boldsymbol{g})_{ij} = \frac{\partial \det(D\boldsymbol{g})}{\partial g_{i,j}}$.

Definition 11.11.2 *Let $\boldsymbol{h}$ be a function defined on an open set, $U \subseteq \mathbb{R}^p$. Then $\boldsymbol{h} \in C^k\left(\overline{U}\right)$ if there exists a function $\boldsymbol{g}$ defined on an open set, W containng $\overline{U}$ such that $\boldsymbol{g} = \boldsymbol{h}$ on U and $\boldsymbol{g}$ is $C^k(W)$.*

Lemma 11.11.3 *There does not exist $\boldsymbol{h} \in C^2\left(\overline{B(\mathbf{0},R)}\right)$ with $\boldsymbol{h} : \overline{B(\mathbf{0},R)} \to \partial B(\mathbf{0},R)$ which has the property that $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{x}$ for all $\boldsymbol{x} \in \partial B(\mathbf{0},R) \equiv \{\boldsymbol{x} : |\boldsymbol{x}| = R\}$ Such a function is called a retract.*

Proof: First note that if $\boldsymbol{h}$ is such a retract, then for all $\boldsymbol{x} \in B(\mathbf{0},R)$,

$$\det(D\boldsymbol{h}(\boldsymbol{x})) = 0$$

This is because if $\det(D\boldsymbol{h}(\boldsymbol{x})) \neq 0$ for some such $\boldsymbol{x}$, then by the inverse function theorem, $\boldsymbol{h}(B(\boldsymbol{x},\delta))$ is an open set for small enough δ but this would require that this open set is a subset of $\partial B(\mathbf{0},R)$ which is impossible because no open ball is contained in $\partial B(\mathbf{0},R)$. Here and below, let B_R denote $\overline{B(\mathbf{0},R)}$.

Now suppose such an $\boldsymbol{h}$ exists. Let $\lambda \in [0,1]$ and let $\boldsymbol{p}_\lambda(\boldsymbol{x}) \equiv \boldsymbol{x} + \lambda(\boldsymbol{h}(\boldsymbol{x}) - \boldsymbol{x})$. This function, $\boldsymbol{p}_\lambda$ is a homotopy of the identity map and the retract $\boldsymbol{h}$. Let

$$I(\lambda) \equiv \int_{B(\mathbf{0},R)} \det(D\boldsymbol{p}_\lambda(\boldsymbol{x}))\,dx.$$

Then using the dominated convergence theorem,

$$\begin{aligned} I'(\lambda) &= \int_{B(\mathbf{0},R)} \sum_{i.j} \frac{\partial \det(D\boldsymbol{p}_\lambda(\boldsymbol{x}))}{\partial p_{\lambda i,j}} \frac{\partial p_{\lambda ij}(\boldsymbol{x})}{\partial \lambda} dx \\ &= \int_{B(\mathbf{0},R)} \sum_i \sum_j \frac{\partial \det(D\boldsymbol{p}_\lambda(\boldsymbol{x}))}{\partial p_{\lambda i,j}} (h_i(\boldsymbol{x}) - x_i)_{,j}\,dx \\ &= \int_{B(\mathbf{0},R)} \sum_i \sum_j \operatorname{cof}(D\boldsymbol{p}_\lambda(\boldsymbol{x}))_{ij} (h_i(\boldsymbol{x}) - x_i)_{,j}\,dx \end{aligned}$$

Now by assumption, $h_i(\boldsymbol{x}) = x_i$ on $\partial B(\mathbf{0},R)$ and so one can integrate by parts, in the iterated integrals used to compute $\int_{B(\mathbf{0},R)}$ and write

$$I'(\lambda) = -\sum_i \int_{B(\mathbf{0},R)} \sum_j \operatorname{cof}(D\boldsymbol{p}_\lambda(\boldsymbol{x}))_{ij,j} (h_i(\boldsymbol{x}) - x_i)\,dx = 0.$$

Therefore, $I(\lambda)$ equals a constant. However, $I(0) = m_p(B(\mathbf{0},R)) \neq 0$ and as pointed out above, $I(1) = 0$. ■

The following is the Brouwer fixed point theorem for C^2 maps.

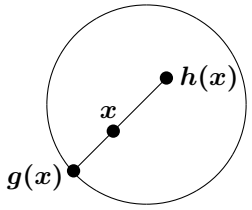
Lemma 11.11.4 *If $\boldsymbol{h} \in C^2\left(\overline{B(\mathbf{0},R)}\right)$ and $\boldsymbol{h} : \overline{B(\mathbf{0},R)} \to \overline{B(\mathbf{0},R)}$, then $\boldsymbol{h}$ has a fixed point, $\boldsymbol{x}$ such that $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{x}$.*

Proof: Suppose the lemma is not true. Then for all $\boldsymbol{x}$, $|\boldsymbol{x} - \boldsymbol{h}(\boldsymbol{x})| \neq 0$. Then define

$$\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{h}(\boldsymbol{x}) + \frac{\boldsymbol{x} - \boldsymbol{h}(\boldsymbol{x})}{|\boldsymbol{x} - \boldsymbol{h}(\boldsymbol{x})|} t(\boldsymbol{x})$$

where $t(\boldsymbol{x})$ is nonnegative and is chosen such that $\boldsymbol{g}(\boldsymbol{x}) \in \partial B(\boldsymbol{0},R)$.

This mapping is illustrated in the following picture.



If $\boldsymbol{x} \to t(\boldsymbol{x})$ is C^2 near $\overline{B(\boldsymbol{0},R)}$, it will follow $\boldsymbol{g}$ is a C^2 retract onto $\partial B(\boldsymbol{0},R)$ contrary to Lemma 11.11.3. Thus $t(\boldsymbol{x})$ is the nonnegative solution t to

$$\left|\boldsymbol{h}(\boldsymbol{x})+\frac{\boldsymbol{x}-\boldsymbol{h}(\boldsymbol{x})}{|\boldsymbol{x}-\boldsymbol{h}(\boldsymbol{x})|}t(\boldsymbol{x})\right|^2=|\boldsymbol{h}(\boldsymbol{x})|^2+2\left(\boldsymbol{h}(\boldsymbol{x}),\frac{\boldsymbol{x}-\boldsymbol{h}(\boldsymbol{x})}{|\boldsymbol{x}-\boldsymbol{h}(\boldsymbol{x})|}\right)t+t^2=R^2 \tag{11.19}$$

then by the quadratic formula,

$$t(\boldsymbol{x})=-\left(\boldsymbol{h}(\boldsymbol{x}),\frac{\boldsymbol{x}-\boldsymbol{h}(\boldsymbol{x})}{|\boldsymbol{x}-\boldsymbol{h}(\boldsymbol{x})|}\right)+\sqrt{\left(\boldsymbol{h}(\boldsymbol{x}),\frac{\boldsymbol{x}-\boldsymbol{h}(\boldsymbol{x})}{|\boldsymbol{x}-\boldsymbol{h}(\boldsymbol{x})|}\right)^2+\left(R^2-|\boldsymbol{h}(\boldsymbol{x})|^2\right)}$$

Is $\boldsymbol{x} \to t(\boldsymbol{x})$ a function in C^2? If what is under the radical is positive, then this is so because $s \to \sqrt{s}$ is smooth for $s > 0$. In fact, this is the case here. The inside of the radical is positive if $R > |\boldsymbol{h}(\boldsymbol{x})|$. If $|\boldsymbol{h}(\boldsymbol{x})| = R$, it is still positive because in this case, the angle between $\boldsymbol{h}(\boldsymbol{x})$ and $\boldsymbol{x}-\boldsymbol{h}(\boldsymbol{x})$ cannot be $\pi/2$. This shows that $\boldsymbol{x} \to t(\boldsymbol{x})$ is the composition of C^2 functions and is therefore C^2. Thus this $\boldsymbol{g}(\boldsymbol{x})$ is a C^2 retract and by the above lemma, there isn't one. ■

Now it is easy to prove the Brouwer fixed point theorem. The following theorem is the Brouwer fixed point theorem for a ball.

Theorem 11.11.5 *Let B_R be the above closed ball and let $\boldsymbol{f}: B_R \to B_R$ be continuous. Then there exists $\boldsymbol{x} \in B_R$ such that $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{x}$.*

Proof: Let $\boldsymbol{f}_k(\boldsymbol{x}) \equiv \frac{\boldsymbol{f}(\boldsymbol{x})}{1+k^{-1}}$. Thus

$$\begin{aligned}\|\boldsymbol{f}_k-\boldsymbol{f}\| &= \max_{\boldsymbol{x}\in B_R}\left\{\left|\frac{\boldsymbol{f}(\boldsymbol{x})}{1+(1/k)}-\boldsymbol{f}(\boldsymbol{x})\right|\right\}=\max_{\boldsymbol{x}\in B_R}\left\{\left|\frac{\boldsymbol{f}(\boldsymbol{x})-\boldsymbol{f}(\boldsymbol{x})(1+(1/k))}{1+(1/k)}\right|\right\}\\ &= \max_{\boldsymbol{x}\in B_R}\left\{\left|\frac{\boldsymbol{f}(\boldsymbol{x})(1/k)}{1+(1/k)}\right|\right\}\le\frac{R}{1+k}\end{aligned}$$

Letting $\|\boldsymbol{h}\| \equiv \max\{|\boldsymbol{h}(\boldsymbol{x})| : \boldsymbol{x} \in B_R\}$, It follows from the Weierstrass approximation theorem, there exists a function whose components are polynomials $\boldsymbol{g}_k$ such that $\|\boldsymbol{g}_k-\boldsymbol{f}_k\| < \frac{R}{k+1}$. Then if $\boldsymbol{x} \in B_R$, it follows

$$\begin{aligned}|\boldsymbol{g}_k(\boldsymbol{x})| &\le |\boldsymbol{g}_k(\boldsymbol{x})-\boldsymbol{f}_k(\boldsymbol{x})|+|\boldsymbol{f}_k(\boldsymbol{x})|\\ &< \frac{R}{1+k}+\frac{kR}{1+k}=R\end{aligned}$$

and so g_k maps B_R to B_R. By Lemma 11.11.4 each of these g_k has a fixed point x_k such that $g_k(x_k) = x_k$. The sequence of points, $\{x_k\}$ is contained in the compact set, B_R and so there exists a convergent subsequence still denoted by $\{x_k\}$ which converges to a point $x \in B_R$. Then from uniform convergence of g_k to f, $f(x) = \lim_{k\to\infty} f(x_k) = \lim_{k\to\infty} g_k(x_k) = \lim_{k\to\infty} x_k = x$. ■

It is not surprising that the ball does not need to be centered at $\mathbf{0}$.

Corollary 11.11.6 *Let $f : \overline{B(a,R)} \to \overline{B(a,R)}$ be continuous. Then there exists $x \in \overline{B(a,R)}$ such that $f(x) = x$.*

Proof: Let $g : B_R \to B_R$ be defined by $g(y) \equiv f(y+a) - a$. Then g is a continuous map from B_R to B_R. Therefore, there exists $y \in B_R$ such that $g(y) = y$. Therefore, $f(y+a) - a = y$ and so letting $x = y + a$, f also has a fixed point as claimed. ■

Definition 11.11.7 *A set A is a retract of a set B if $A \subseteq B$, and there is a continuous map $h : B \to A$ such that $h(x) = x$ for all $x \in A$ and h is onto. B has the fixed point property means that whenever g is continuous and $g : B \to B$, it follows that g has a fixed point.*

Proposition 11.11.8 *Let A be a retract of B and suppose B has the fixed point property. Then so does A.*

Proof: Suppose $f : A \to A$. Let h be the retract of B onto A. Then $f \circ h : B \to B$ is continuous. Thus, it has a fixed point $x \in B$ so $f(h(x)) = x$. However, $h(x) \in A$ and $f : A \to A$ so in fact, $x \in A$. Now $h(x) = x$ and so $f(x) = x$. ■

Recall that every convex compact subset K of $\mathbb{R}^p$ is a retract of all of $\mathbb{R}^p$ obtained by using the projection map. See Problems beginning with 8 on Page 152. In particular, K is a retract of a large closed ball containing K, which ball has the fixed point property. Therefore, K also has the fixed point property. This shows the following which is a convenient formulation of the Brouwer fixed point theorem. However, Proposition 11.11.8 is more general.

Theorem 11.11.9 *Every convex closed and bounded subset of $\mathbb{R}^p$ has the fixed point property.*

11.12 Exercises

1. Suppose $A \subseteq \mathbb{R}^p$ is covered by a finite collection of Balls $\mathscr{F}$. Show that then there exists a disjoint collection of these balls, $\{B_i\}_{i=1}^m$, such that $A \subseteq \cup_{i=1}^m \widehat{B}_i$ where $\widehat{B}_i$ has the same center as B_i but 3 times the radius. **Hint:** Since the collection of balls is finite, they can be arranged in order of decreasing radius. Mimic the argument for Vitali covering theorem.

2. Show that if f is absolutely continuous on $[a,b]$ and if $V(x)$ is the total variation of f on $[0,x]$, then V is also absolutely continuous.

3. In Problem 5 on Page 315, you showed that if $f \in L^1(\mathbb{R}^p)$, there exists h which is continuous and equal to 0 off some compact set such that $\int |f-h|\,dm < \varepsilon$. Define $f_y(x) \equiv f(x-y)$. Explain why f_y is Lebesgue measurable and $\int |f_y|\,dm_p = \int |f|\,dm_p$. Now justify the following formula. $\int |f_y - f|\,dm_p \leq \int |f_y - h_y|\,dm_p +$

$\int |h_{\boldsymbol{y}} - h|\, dm_p + \int |h - f|\, dm_p \leq 2\varepsilon + \int |h_{\boldsymbol{y}} - h|\, dm_p$. Now explain why the last term is less than ε if $\|\boldsymbol{y}\|$ is small enough. Explain continuity of translation in $L^1(\mathbb{R}^p)$ which says that $\lim_{\boldsymbol{y}\to \mathbf{0}} \int_{\mathbb{R}^p} |f_{\boldsymbol{y}} - f|\, dm_p = 0$

4. This problem will help to understand that a certain kind of function exists. Let $f(x) = e^{-1/x^2}$ if $x \neq 0$ and let $f(x) = 0$ if $x = 0$. Show that f is infinitely differentiable. Note that you only need to be concerned with what happens at 0. There is no question elsewhere. This is a little fussy but is not too hard.

5. ↑Let $f(x)$ be as given above. Now let $\hat{f}(x) = f(x)$ if $x \leq 0$ and let $\hat{f}(x) = 0$ if $x > 0$. Show that $\hat{f}(x)$ is also infinitely differentiable. Now let $r > 0$ and define $g(x) \equiv \hat{f}(-(x-r))\,\hat{f}(x+r)$. Show that g is infinitely differentiable and vanishes for $|x| \geq r$. Let $\psi(\boldsymbol{x}) = \prod_{k=1}^{p} g(x_k)$. For $U = B(\mathbf{0}, 2r)$ with the norm given by $\|\boldsymbol{x}\| = \max\{|x_k|, k \leq p\}$, show that $\psi \in C_c^\infty(U)$.

6. ↑Using the above problem, show there exists $\psi \geq 0$ such that $\psi \in C_c^\infty(B(\mathbf{0}, 1))$ and $\int \psi dm_p = 1$. Now define $\psi_n(\boldsymbol{x}) \equiv n^p \psi(n\boldsymbol{x})$. Show that ψ_n equals zero off a compact subset of $B\left(\mathbf{0}, \frac{1}{n}\right)$ and $\int \psi_n dm_p = 1$. We say that $\operatorname{spt}(\psi_n) \subseteq B\left(\mathbf{0}, \frac{1}{n}\right)$. $\operatorname{spt}(f)$ is defined as the closure of the set on which f is not equal to 0. Such a sequence of functions as just defined $\{\psi_n\}$ where $\int \psi_n dm_p = 1$ and $\psi_n \geq 0$ and $\operatorname{spt}(\psi_n) \subseteq B\left(\mathbf{0}, \frac{1}{n}\right)$ is called a **mollifier.**

7. ↑It is important to be able to approximate functions with those which are infinitely differentiable. Suppose $f \in L^1(\mathbb{R}^p)$ and let $\{\psi_n\}$ be a mollifier as above. We define the convolution as follows.

$$f * \psi_n(\boldsymbol{x}) \equiv \int f(\boldsymbol{x} - \boldsymbol{y})\, \psi_n(\boldsymbol{y})\, dm_p(\boldsymbol{y})$$

Here the notation means that the variable of integration is $\boldsymbol{y}$. Show that $f * \psi_n(\boldsymbol{x})$ exists and equals $\int \psi_n(\boldsymbol{x} - \boldsymbol{y})\, f(\boldsymbol{y})\, dm_p(\boldsymbol{y})$. Now show using the dominated convergence theorem that $f * \psi_n$ is infinitely differentiable. Next show that

$$\lim_{n\to\infty} \int |f(\boldsymbol{x}) - f * \psi_n(\boldsymbol{x})|\, dm_p = 0.$$

Thus, in terms of being close in $L^1(\mathbb{R}^p)$, every function in $L^1(\mathbb{R}^p)$ is close to one which is infinitely differentiable.

8. ↑From Problem 5 above and $f \in L^1(\mathbb{R}^p)$, there exists $h \in C_c(\mathbb{R}^p)$, continuous and $\operatorname{spt}(h)$ a compact set, such that $\int |f - h|\, dm_p < \varepsilon$. Now consider $h * \psi_n$. Show that this function is in $C_c^\infty\left(\operatorname{spt}(h) + B\left(\mathbf{0}, \frac{2}{n}\right)\right)$. The notation means you start with the compact set $\operatorname{spt}(h)$ and fatten it up by adding the set $B\left(\mathbf{0}, \frac{1}{n}\right)$. It means $\boldsymbol{x} + \boldsymbol{y}$ such that $\boldsymbol{x} \in \operatorname{spt}(h)$ and $\boldsymbol{y} \in B\left(\mathbf{0}, \frac{1}{n}\right)$. Show the following. For all n large enough, $\int |f - h * \psi_n|\, dm_p < \varepsilon$ so one can approximate with a function which is infinitely differentiable and also has compact support. Also show that $h * \psi_n$ converges uniformly to h. If h is a function in $C^k(\mathbb{R}^n)$ in addition to being continuous with compact support, show that for each $|\alpha| \leq k, D^\alpha(h * \psi_n) \to D^\alpha h$ uniformly. **Hint:** If you do this for a single partial derivative, you will see how it works in general.

9. ↑Let $f \in L^1(\mathbb{R})$. Show that $\lim_{n\to\infty} \int f(x)\sin(nx)\,dm = 0$. **Hint:** Use the result of the above problem to obtain $g \in C_c^\infty(\mathbb{R})$, continuous and zero off a compact set, such that $\int |f-g|\,dm < \varepsilon$. Then show that

$$\lim_{n\to\infty} \int g(x)\sin(nx)\,dm(x) = 0$$

You can do this by integration by parts. Then consider this. $|\int f(x)\sin(nx)\,dm| =$

$$\left|\int f(x)\sin(nx)\,dm - \int g(x)\sin(nx)\,dm\right| + \left|\int g(x)\sin(nx)\,dm\right|$$

$$\leq \int |f-g|\,dm + \left|\int g(x)\sin(nx)\,dm\right|$$

This is the celebrated Riemann Lebesgue lemma which is the basis for all theorems about pointwise convergence of Fourier series and Fourier integrals.

10. As another application of theory of regularity, here is a very important result. Suppose $f \in L^1(\mathbb{R}^p)$ and for every $\psi \in C_c^\infty(\mathbb{R}^p)$ $\int f\psi dm_p = 0$. Show that then it follows $f(\boldsymbol{x}) = 0$ for $a.e.\boldsymbol{x}$. That is, there is a set of measure zero such that off this set f equals 0. **Hint:** What you can do is to let E be a measurable set which is bounded and let $K_n \subseteq E \subseteq V_n$ where $m_p(V_n \setminus K_n) < 2^{-n}$. Here K_n is compact and V_n is open. By an earlier exercise, Problem 11 on Page 255, there exists a function ϕ_n which is continuous, has values in $[0,1]$ equals 1 on K_n and $\mathrm{spt}(\phi_n) \subseteq V$. To get this last part, show there exists W_n open such that $\bar{W}_n \subseteq V_n$ and W_n contains K_n. Then you use the problem to get $\mathrm{spt}(\phi_n) \subseteq \bar{W}_n$. Now you form $\eta_n = \phi_n * \psi_l$ where $\{\psi_l\}$ is a mollifier. Show that for l large enough, η_n has values in $[0,1]$, $\mathrm{spt}(\eta_n) \subseteq V_n$ and $\eta_n \in C_c^\infty(V_n)$. Now explain why $\eta_n \to \mathcal{X}_E$ off a set of measure zero. To do this, you might want to consider the Borel Cantelli lemma, Lemma 9.2.5 on Page 247. Then

$$\left|\int f\mathcal{X}_E dm_p\right| = \left|\int f(\mathcal{X}_E - \eta_n)\,dm_p\right| + \left|\int f\eta_n dm_p\right| = \left|\int f(\mathcal{X}_E - \eta_n)\,dm_p\right|$$

Now explain why this converges to 0 on the right. This will involve the dominated convergence theorem. Conclude that $\int f\mathcal{X}_E dm_p = 0$ for every bounded measurable set E. Show that this implies that $\int f\mathcal{X}_E dm_p = 0$ for every measurable E. Explain why this requires $f = 0$ a.e.

11. Suppose f, g are absolutely continuous on $[a,b]$. Prove the usual integration by parts formula. **Hint:** You might try the following:

$$\int_a^b f'(t)\,g(t)\,dm(t) = \int_a^b f'(t)\left(\int_a^t g'(s)\,dm(s) + g(a)\right)dm(t)$$

Now do one integration and then interchange the order of integration.

12. Let $x \to F(f(x))$ be absolutely continuous whenever $x \to f(x)$ is absolutely continuous. Then F is Lipschitz. This is due to G. M. Fishtenholz. **Hint:** Reduce to Proposition 11.7.5 using an appropriate Lipschitz continuous function f.

13. Suppose $g:[c,d]\to[a,b]$ and is absolutely continuous and increasing and $f:[a,b]\to\mathbb{R}$ is Lipschitz continuous. Show that then $f\circ g$ is absolutely continuous and

$$\int_a^b f(t)\,dm(t)=\int_c^d f'(g(s))\,g'(s)\,ds=f(b)-f(a)$$

14. If $f\in L^1(\Omega,\mu)$, show that $\lim_{\mu(E)\to 0}\int_E |f|\,d\mu=0$. Defining $F(x)\equiv\int_a^x f(t)\,dm(t)$ for $f\in L^1([a,b],m)$, verify that F is absolutely continuous with $F'(x)=f(x)$ a.e.

15. Show that if f is absolutely continuous, as defined in Definition 11.7.2, then it is of bounded variation.

16. Let $f:[a,b]\to\mathbb{R}$ be absolutely continuous. Show that in fact, the total variation of f on $[a,b]$ is $\int_a^b |f'|\,dm$. **Hint:** One direction is easy, that $V[a,b]\le\int_a^b |f'|\,dm$. To do the other direction, show there is a sequence of **step functions** $s_n(t)\equiv\sum_{k=1}^{m_n}\alpha_k^n\mathcal{X}_{I_k^n}(t)$ which converges to $\operatorname{sgn} f'$ pointwise a.e., $I_k^n=\left(c_{k-1}^n,c_k^n\right)$. This will involve regularity notions. Explain why it can be assumed each $\left|\alpha_k^n\right|\le 1$. Then

$$\left|\int_a^b f's_n\right|=\left|\sum_{k=1}^{m_n}\alpha_k^n\int_{I_k^n}f'\right|\le\sum_{k=1}^{m_n}\left|f(c_k^n)-f\left(c_{k-1}^n\right)\right|\le V([a,b],f)$$

Now pass to a limit using the dominated convergence theorem.

17. Let $F(x)=\left(\int_0^x e^{-t^2}dt\right)^2$. Justify the following:

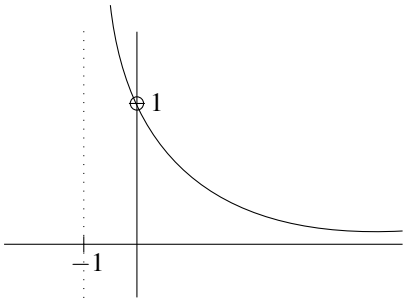
$$F'(x)=2\left(\int_0^x e^{-t^2}dt\right)e^{-x^2}=2xe^{-x^2}\left(\int_0^1 e^{-x^2t^2}dt\right)=2x\left(\int_0^1 e^{-x^2\left(t^2+1\right)}dt\right)$$

Now integrate.

$$\begin{aligned}F(x)&=\int_0^x\int_0^1 2ue^{-u^2\left(t^2+1\right)}dtdu=\int_0^1\int_0^x 2ue^{-u^2\left(t^2+1\right)}dudt\\&=\int_0^1 -e^{-u^2\left(t^2+1\right)}\frac{1}{1+t^2}\Big|_0^x dt=\int_0^1\left(\frac{1}{1+t^2}-e^{-x^2}\frac{1}{1+t^2}\right)dt\end{aligned}$$

Now let $x\to\infty$ and conclude $F(\infty)=\left(\int_0^\infty e^{-t^2}dt\right)^2=\int_0^1\frac{1}{1+t^2}dt=\frac{\pi}{4}$.

18. This problem outlines an approach to Stirling's formula following [48] and [7]. From the above problems, $\Gamma(n+1)=n!$ for $n\ge 0$. Consider more generally $\Gamma(x+1)$ for $x>0$. Actually, we will always assume $x>1$ since it is the limit as $x\to\infty$ which is of interest. $\Gamma(x+1)=\int_0^\infty e^{-t}t^x dt$. Change variables letting $t=x(1+u)$ to obtain $\Gamma(x+1)=x^{x+1}e^{-x}\int_{-1}^\infty\left((1+u)e^{-u}\right)^x du$. Next let $h(u)$ be such that $h(0)=1$ and $(1+u)e^{-u}=\exp\left(-\frac{u^2}{2}h(u)\right)$. Show that the thing which works is $h(u)=\frac{2}{u^2}(u-\ln(1+u))$. Use L'Hospital's rule to verify that the limit of $h(u)$ as $u\to 0$ is 1. The graph of h is illustrated in the following picture. Verify that its graph is like this, with an asymptote at $u=-1$ decreasing and equal to 1 at 0 and converging to 0 as $u\to\infty$.



Next change the variables again letting $u = s\sqrt{\frac{2}{x}}$. This yields, from the original description of h

$$\Gamma(x+1) = x^x e^{-x}\sqrt{2x}\int_{-\sqrt{x/2}}^{\infty} \exp\left(-s^2 h\left(s\sqrt{\frac{2}{x}}\right)\right) ds$$

For $s < 1, h\left(s\sqrt{\frac{2}{x}}\right) > 2 - 2\ln 2 = 0.613\,71$ so the above integrand is dominated by $e^{-(2-2\ln 2)s^2}$. Consider the integrand in the above for $s > 1$. Show that the exponent part is

$$-\left(\sqrt{2}\sqrt{x}s - x\ln\left(1+s\sqrt{\frac{2}{x}}\right)\right)$$

The expression $\left(\sqrt{2}\sqrt{x}s - x\ln\left(1+s\sqrt{\frac{2}{x}}\right)\right)$ is increasing in x. You can show this by fixing s and taking a derivative with respect to x. Therefore, it is larger than $\sqrt{2}\sqrt{1}s - \ln\left(1+s\sqrt{\frac{2}{1}}\right)$ and so

$$\begin{aligned}\exp\left(-s^2 h\left(s\sqrt{\frac{2}{x}}\right)\right) &\leq \exp\left(-\left(\sqrt{2}\sqrt{1}s - \ln\left(1+s\sqrt{\frac{2}{1}}\right)\right)\right)\\ &= \left(1+s\sqrt{2}\right)e^{-\sqrt{2}s}\end{aligned}$$

Thus, there exists a dominating function for $\mathscr{X}_{\left[-\sqrt{\frac{x}{2}},\infty\right]}(s)\exp\left(-s^2 h\left(s\sqrt{\frac{2}{x}}\right)\right)$ and these functions converge pointwise to $\exp\left(-s^2\right)$ as $x \to \infty$ so by the dominated convergence theorem,

$$\lim_{x\to\infty}\int_{-\sqrt{x/2}}^{\infty}\exp\left(-s^2 h\left(s\sqrt{\frac{2}{x}}\right)\right) ds = \int_{-\infty}^{\infty} e^{-s^2} ds = \sqrt{\pi}$$

See Problem 17. This yields a general Stirling's formula, $\lim_{x\to\infty}\frac{\Gamma(x+1)}{x^x e^{-x}\sqrt{2x}} = \sqrt{\pi}$.

19. Let $\bar{\boldsymbol{\mu}}$ be the completion of the measure $\boldsymbol{\mu} = \prod_{i=1}^p \mu_i$ on the product σ algebra $\prod_{i=1}^p \mathscr{F}_i$. Show that if $f \geq 0$ is $\overline{\prod_{i=1}^p \mathscr{F}_i}$ measurable, then there are two functions g, h which are $\prod_{i=1}^p \mathscr{F}_i$ measurable and $g \geq f \geq h$ while $\boldsymbol{\mu}([g-h>0]) = 0$. Assume the original measure spaces are finite or σ finite.

20. In the representation theorem for positive linear functionals, show that if $\mu(V) = \sup\{Lf : f \prec V\}$, then the σ algebra and measure representing the functional are unique.

11.13 Change of Variables, Linear Maps

The change of variables formula for linear maps is implied by the following major theorem. First, here is some notation. Let $\boldsymbol{dx}_i$ denote $dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_p$.

Next is an important observation about measurability and Lipschitz maps on $\mathbb{R}^p$. A nice property of Lipschitz functions ($|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{y})| \leq K|\boldsymbol{x}-\boldsymbol{y}|$) is that they take Lebesgue measurable sets to Lebesgue measurable sets. I will base the following argument on Proposition 4.5.3, the Vitali covering theorem. It is possible to use the Besicovitch covering theorem also and you might try to do this. When you are dealing with Lebesgue measure rather than some general Radon measure of the sort you encounter in probability, I think the Vitali theorem tends to be a little easier. The difficulty with using it in general is that for a general Radon measure, if you modify the radius, you end up not knowing much about the measure of the resulting ball. However, if you have a ball in $\mathbb{R}^p$ $B(\boldsymbol{x},r)$, then $m(B(\boldsymbol{x},\alpha r)) = \alpha^p m(B(\boldsymbol{x},r))$. This is most easily seen if the ball is taken with respect to $\|\boldsymbol{x}\| \equiv \max\{|x_i|, i \leq p\}$. In this case you have a p dimensional cube with $\boldsymbol{x}$ at the center and so that which was just claimed is obvious because the measure of the "ball" is the product of the lengths of the sides. In what follows, you might want to simply assume this is the way distance is measured. It will end up not mattering at all, which follows from the change of variables results to be presented, but such claims are more obvious with this assumption.

Theorem 11.13.1 *In case $\boldsymbol{h}$ is Lipschitz, $\|\boldsymbol{h}(\boldsymbol{x}) - \boldsymbol{h}(\boldsymbol{y})\| \leq K\|\boldsymbol{x}-\boldsymbol{y}\|$ then if T is a set for which $m_p(T) = 0$, it follows that $m_p(\boldsymbol{h}(T)) = 0$. Also if $E \in \mathscr{F}_p$, then $\boldsymbol{h}(E) \in \mathscr{F}_p$.*

Proof: By the Lipschitz condition, $\|\boldsymbol{h}(\boldsymbol{x}+\boldsymbol{v}) - \boldsymbol{h}(\boldsymbol{x})\| \leq K\|\boldsymbol{v}\|$ and you can simply let $T \subseteq V$ where $m_p(V) < \varepsilon/(K^p 5^p)$. Then there is a countable disjoint sequence of balls $\{B_i\}$ such that $\{\hat{B}_i\}$ covers T and each ball B_i is contained in V each having radius no more than 1. Then $\boldsymbol{h}(\hat{B}_i) \subseteq B(\boldsymbol{h}(\boldsymbol{x}_i), 5K)$ and so

$$\bar{m}_p(\boldsymbol{h}(T)) \leq \sum_{i=1}^{\infty} m_p(\boldsymbol{h}(\hat{B}_i)) \leq 5^p K^p \sum_{i=1}^{\infty} m_p(B_i) \leq K^p 5^p m_p(V) < \varepsilon$$

Since ε is arbitrary, this shows that $\boldsymbol{h}(T)$ is measurable and $m_p(\boldsymbol{h}(T)) = 0$.

Now let $E \in \mathscr{F}_p$, $m_p(E) < \infty$. Then by regularity, there exists F which is the countable union of compact sets such that $E = F \cup N$ where N is a set of measure zero. Then from the first part, $\boldsymbol{h}(E \setminus F) \subseteq \boldsymbol{h}(N)$ and this set on the right has measure zero and so by completeness of the measure, $\boldsymbol{h}(E \setminus F) \in \mathscr{F}_p$ and so

$$\boldsymbol{h}(E) = \boldsymbol{h}(E \setminus F) + \boldsymbol{h}(F) \in \mathscr{F}_p$$

because $F = \cup_k K_k$, each K_k compact. Hence $\boldsymbol{h}(F) = \cup_k \boldsymbol{h}(K_k)$ which is the countable union of compact sets due to the continuity of $\boldsymbol{h}$. For arbitrary E, $\boldsymbol{h}(E) = \cup_{k=1}^{\infty} \boldsymbol{h}(E \cap B(\boldsymbol{0},k)) \in \mathscr{F}_p$. ■

Of course an example of a Lipschitz map is a linear map. $\|A\boldsymbol{x} - A\boldsymbol{y}\| = \|A(\boldsymbol{x}-\boldsymbol{y})\| \leq \|A\|\|\boldsymbol{x}-\boldsymbol{y}\|$. Therefore, if A is linear and E is Lebesgue measurable, then AE is also Lebesgue measurable. This is convenient.

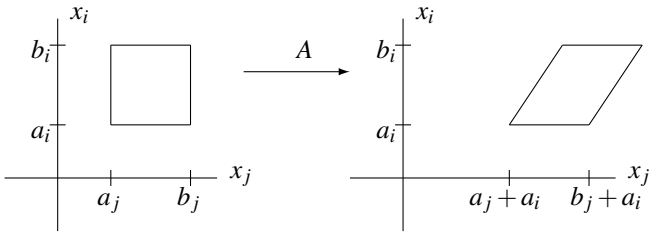
Theorem 11.13.2 *Let E be any Lebesgue measurable set and let A be a $p \times p$ matrix. Then AE is Lebesgue measurable and $m_p(AE) = |\det(A)| m_p(E)$. Also, if E is any Lebesgue measurable set, then*

$$\int \mathcal{X}_{A(E)}(\boldsymbol{y})\, dy = \int \mathcal{X}_E(\boldsymbol{x}) |\det(A)|\, dx.$$

Proof: Let Q denote $\prod_{i=1}^p (a_i, b_i)$, an open rectangle. First suppose A is an elementary matrix which is obtained from I by adding α times row i to row j. Then $\det(A) = 1$ and

$$\begin{aligned} m_p(AQ) &\equiv \left(\prod_{i\neq j}\int_{a_i}^{b_i}\right)\int \mathcal{X}_{AQ} dx_j d\boldsymbol{x}_j = \left(\prod_{i\neq j}\int_{a_i}^{b_i}\right)\int_{a_j+\alpha x_i}^{b_j+\alpha x_i} dx_j d\boldsymbol{x}_j \\ &= \prod_{i=1}^p |b_i - a_i| = m_p(Q) = |\det(A)| m_p(Q) \end{aligned}$$

The linear transformation determined by A just shears Q in the direction of x_j. It is clear in this case that AQ is an open, hence measurable set. A picture of the effect of doing A to Q follows in case $\alpha = 1$.



Next suppose A is an elementary matrix which comes from multiplying the j^{th} row of I with $\alpha \neq 0$. Then this changes the length of one of the sides of the box by the factor α, resulting in

$$m_p(AQ) = |\alpha| |b_j - a_j| \prod_{i\neq j} |b_i - a_i| = |\alpha| m_p(Q) = |\det(A)| m_p(Q)$$

In case A is an elementary matrix which comes from switching two rows of the identity, then it maps Q to a possibly different Q which has the lengths of its sides the same set of numbers as the set of lengths of the sides of Q. Therefore, $|\det(A)| = 1$ and so

$$m_p(AQ) = m_p(Q) = |\det(A)| m_p(Q).$$

Let $R_n = \prod_{i=1}^p (-n, n)$ and let $\mathcal{K}$ consist of all finite intersections of such open boxes as just described. This is a π system. Now let

$$\mathcal{G} \equiv \left\{E \in \mathcal{B}(\mathbb{R}^p) : m_p(A(E \cap R_n)) = |\det(A)| m_p(E \cap R_n), \text{ and } AE \text{ is Borel}\right\}$$

where A is any of the above elementary matrices. It is clear that $\mathcal{G}$ is closed with respect to countable disjoint unions. If $E \in \mathcal{G}$, is $E^C \in \mathcal{G}$? Since A is onto, $AE^C = \mathbb{R}^n \setminus AE$ which is Borel. What about the estimate?

$$m_p(A(E \cap R_n)) + m_p\left(A\left(E^C \cap R_n\right)\right) = m_p(A(R_n)) = |\det(A)| m_p(R_n)$$

and so

$$\begin{aligned} m_p\left(A\left(E^C \cap R_n\right)\right) &= |\det(A)|\, m_p(R_n) - m_p(A(E \cap R_n)) \\ &= |\det(A)|\,(m_p(R_n) - m_p(E \cap R_n)) = |\det(A)|\, m_p\left(E^C \cap R_n\right) \end{aligned}$$

From above, $\mathscr{G}$ contains $\mathscr{K}$. By Lemma 9.3.2, it follows that $\mathscr{G} \supseteq \sigma(\mathscr{K}) = \mathscr{B}(\mathbb{R}^p)$. Therefore, for any A elementary and E a Borel set,

$$|\det(A)|\, m_p(E) = \lim_{n\to\infty} |\det(A)|\, m_p(E \cap R_n) = \lim_{n\to\infty} m_p(A(E \cap R_n)) = m_p(A(E)).$$

Now consider A an arbitrary invertible matrix. Then A is the product of elementary matrices $A_1 \cdots A_m$ and so if E is Borel, so is AE. Also

$$|\det(A)|\, m_p(E) = \prod_{i=1}^{m} |\det(A_i)|\, m_p(E) = \prod_{i=1}^{m-1} |\det(A_i)|\, m_p(A_m E)$$

$$= \prod_{i=1}^{m-2} |\det(A_i)|\, m_p(A_{m-1}A_m E) \cdots = m_p(A_1 \cdots A_m E) = m_p(AE).$$

In case A is an arbitrary $p \times p$ matrix which has rank less than p, there exists a sequence of elementary matrices $E_1, E_2, \cdots, E_s$ such that $A = E_1 E_2 \cdots E_s B$ where B is in row reduced echelon form and has at least one row of zeros. Thus if S is any Lebesgue measurable set, $A(S) = E_1 E_2 \cdots E_s B(S)$ and so

$$\begin{aligned} m_p(A(S)) &= m_p(E_1 E_2 \cdots E_s B(S)) \le m_p(E_1 E_2 \cdots E_s B(\mathbb{R}^p)) \\ &= |\det(E_1 E_2 \cdots E_s)|\, m_p(B(\mathbb{R}^p)) \end{aligned}$$

However, $B(\mathbb{R}^n)$ is a measurable set of measure zero because it is contained in a set of the form $F \equiv \left\{\boldsymbol{x} \in \mathbb{R}^p : x_k, \cdots, x_p = 0\right\}$. and this has measure zero.

It has now been shown that for invertible A, and E any Borel set, it follows that $m_p(AE) = |\det(A)|\, m_p(E)$ and for any Lebesgue measurable set E and A not invertible, the above formula holds. It only remains to verify the formula holds for A invertible and E only Lebesgue measurable.

However, in this case, A maps open sets to open sets because its inverse is continuous and maps compact sets to compact sets because $\boldsymbol{x} \to A\boldsymbol{x}$ is continuous. Hence A takes G_δ sets to G_δ sets and F_σ sets to F_σ sets. Let E be bounded and Lebesgue measurable. By regularity of the measure, there exists G and F, G_δ and F_σ sets respectively such that $F \subseteq E \subseteq G$ and $m_p(G) = m_p(E) = m_p(F)$. Then $AF \subseteq AE \subseteq AG$ and and for $\bar{m}_p$ the outer measure determined by m_p,

$$\begin{aligned} |\det(A)|\, m_p(F) &= m_p(AF) \le \bar{m}_p(AE) \le m_p(AG) \\ &= \det(A)\, m_p(G) = |\det(A)|\, m_p(E) = |\det(A)|\, m_p(F) \end{aligned}$$

Thus all inequalities are equal signs. By completeness, AE is Lebesgue measurable because

$$0 = m_p(AG \setminus AF) = |\det(A)|\, m_p(G \setminus F) \ge \bar{m}_p(AE \setminus AF)$$

For arbitrary E, let $E_n \equiv E \cap B(\mathbf{0}, n)$. Then AE_n is measurable and so the union of these AE_n which is AE is also measurable and

$$m_p(AE) = \lim_{n\to\infty} m_p(AE_n) = |\det(A)| \lim_{n\to\infty} m_p(E_n) = |\det(A)|\, m_p(E) \ \blacksquare$$

Theorem 11.13.3 *Let $f \geq 0$ and suppose it is Lebesgue measurable. Then if A is a $p \times p$ matrix,*

$$\int \mathcal{X}_{A(\mathbb{R}^p)}(\boldsymbol{y}) f(\boldsymbol{y})\, dy = \int f(A(\boldsymbol{x}))\, |\det(A)|\, dx. \tag{11.20}$$

Proof: From Theorem 11.13.2, the equation is true if $\det(A) = 0$. It follows that it suffices to consider only the case where A^{-1} exists. First suppose $f(\boldsymbol{y}) = \mathcal{X}_E(\boldsymbol{y})$ where E is a Lebesgue measurable set. In this case, $A(\mathbb{R}^n) = \mathbb{R}^n$. Then from Theorem 11.13.2

$$\int \mathcal{X}_{A(\mathbb{R}^p)}(\boldsymbol{y}) f(\boldsymbol{y})\, dy = \int \mathcal{X}_E(\boldsymbol{y})\, dy = m_p(E) = |\det(A)|\, m_p\left(A^{-1}E\right)$$

$$= \int_{\mathbb{R}^n} |\det(A)|\, \mathcal{X}_{A^{-1}E}(\boldsymbol{x})\, dx = \int_{\mathbb{R}^n} |\det(A)|\, \mathcal{X}_E(A\boldsymbol{x})\, dx = \int f(A(\boldsymbol{x}))\, |\det(A)|\, dx$$

It follows from this that 11.20 holds whenever $\boldsymbol{f}$ is a nonnegative simple function. Finally, the general result follows from approximating the Lebesgue measurable function with nonnegative simple functions using Theorem 9.1.6 and then applying the monotone convergence theorem. ■

Note that it now follows that for the ball defined by any norm,

$$m_p(B(\boldsymbol{x}, \alpha r)) = |\alpha|^p\, m_p(B(\boldsymbol{x}, r))$$

because the linear transformation $\boldsymbol{x} \to \alpha\boldsymbol{x}$ has determinant equal to α^p.

This is now a very good change of variables formula for a linear transformation. Next this is extended to differentiable functions.

11.14 Differentiable Functions and Measurability

To begin with, certain kinds of functions map measurable sets to measurable sets. It was shown earlier, Theorem 11.13.1, that Lipschitz functions do this. So do differentiable functions.

As pointed out earlier, it is convenient to take all balls with respect to the norm on $\mathbb{R}^p$ given by $\|\boldsymbol{x}\| = \max\{|x_k| : k = 1, 2, \cdots, p\}$. Thus from the definition of this norm, $B(\boldsymbol{x}, r)$ is the open box, $\prod_{k=1}^p (x_k - r, x_k + r)$ and so $m_p(B(\boldsymbol{x}, r)) = (2r)^p = 2^p r^p$. Also for a linear transformation $A \in \mathcal{L}(\mathbb{R}^p, \mathbb{R}^p)$, I will continue to use $\|A\| \equiv \sup_{\|\boldsymbol{x}\| \leq 1} \|A\boldsymbol{x}\|$.

Lemma 11.14.1 *Let $T \subseteq U$, where U is open, $\boldsymbol{h}$ is continuous, and let $\boldsymbol{h}$ be differentiable at each $\boldsymbol{x} \in T$ and suppose that $m_p(T) = 0$, then $m_p(\boldsymbol{h}(T)) = 0$.*

Proof: For $k \in \mathbb{N}$, $T_k \equiv \{\boldsymbol{x} \in T : \|D\boldsymbol{h}(\boldsymbol{x})\| < k\}$ and let $\varepsilon > 0$ be given. Since T_k is a subset of a set of measure zero, it is measurable, but we don't need to pay much attention to this fact. Now by outer regularity, there exists an open set V, containing T_k which is contained in U such that $m_p(V) < \varepsilon$. Let $\boldsymbol{x} \in T_k$. Then by differentiability, $\boldsymbol{h}(\boldsymbol{x} + \boldsymbol{v}) = \boldsymbol{h}(\boldsymbol{x}) + D\boldsymbol{h}(\boldsymbol{x})\boldsymbol{v} + \boldsymbol{o}(\boldsymbol{v})$ and so there exist arbitrarily small $r_{\boldsymbol{x}} < 1$ such that $B(\boldsymbol{x}, 5r_{\boldsymbol{x}}) \subseteq V$ and whenever $\|\boldsymbol{v}\| \leq 5r_{\boldsymbol{x}}$, $\|\boldsymbol{o}(\boldsymbol{v})\| < \frac{1}{5}\|\boldsymbol{v}\|$. Thus, from the Vitali covering theorem, Theorem 4.5.3,

$$\begin{aligned} \boldsymbol{h}(B(\boldsymbol{x}, 5r_{\boldsymbol{x}})) &\subseteq D\boldsymbol{h}(\boldsymbol{x})(B(\boldsymbol{0}, 5r_{\boldsymbol{x}})) + \boldsymbol{h}(\boldsymbol{x}) + B(\boldsymbol{0}, r_{\boldsymbol{x}}) \subseteq B(\boldsymbol{0}, k5r_{\boldsymbol{x}}) + \\ +B(\boldsymbol{0}, r_{\boldsymbol{x}}) + \boldsymbol{h}(\boldsymbol{x}) &\subseteq B(\boldsymbol{h}(\boldsymbol{x}), (5k+1)\, r_x) \subseteq B(\boldsymbol{h}(\boldsymbol{x}), 6kr_{\boldsymbol{x}}) \end{aligned}$$

From the Vitali covering theorem, there exists a countable disjoint sequence of these balls, $\{B(\boldsymbol{x}_i, r_i)\}_{i=1}^{\infty}$ such that $\{B(\boldsymbol{x}_i, 5r_i)\}_{i=1}^{\infty} = \left\{\widehat{B}_i\right\}_{i=1}^{\infty}$ covers T_k. Then letting $\overline{m_p}$ denote the outer measure determined by m_p,

$$\overline{m_p}(\boldsymbol{h}(T_k)) \leq \overline{m_p}\left(\boldsymbol{h}\left(\cup_{i=1}^{\infty}\widehat{B}_i\right)\right) \leq \sum_{i=1}^{\infty}\overline{m_p}\left(\boldsymbol{h}\left(\widehat{B}_i\right)\right)$$

$$\leq \sum_{i=1}^{\infty} m_p(B(\boldsymbol{h}(\boldsymbol{x}_i), 6kr_{\boldsymbol{x}_i})) = \sum_{i=1}^{\infty} m_p(B(\boldsymbol{x}_i, 6kr_{\boldsymbol{x}_i}))$$

$$= (6k)^p \sum_{i=1}^{\infty} m_p(B(\boldsymbol{x}_i, r_{\boldsymbol{x}_i})) \leq (6k)^p m_p(V) \leq (6k)^p \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, this shows $m_p(\boldsymbol{h}(T_k)) = \overline{m_p}(\boldsymbol{h}(T_k)) = 0$. Now $m_p(\boldsymbol{h}(T)) = \lim_{k\to\infty} m_p(\boldsymbol{h}(T_k)) = 0$. ■

Lemma 11.14.2 *Let $\boldsymbol{h}$ be continuous on U and let $\boldsymbol{h}$ be differentiable on $T \subseteq U$. If S is a Lebesgue measurable subset of T, then $\boldsymbol{h}(S)$ is Lebesgue measurable.*

Proof: By Theorem 11.9.2 there exists F which is a countable union of compact sets $F = \cup_{k=1}^{\infty} K_k$ such that $F \subseteq S$, $m_p(S \setminus F) = 0$.Then $\boldsymbol{h}(F) = \cup_k \boldsymbol{h}(K_k) \in \mathscr{B}(\mathbb{R}^p)$ because the continuous image of a compact set is compact. Also, $\boldsymbol{h}(S \setminus F)$ is a set of measure zero by Lemma 11.14.1 and so $\boldsymbol{h}(S) = \boldsymbol{h}(F) \cup \boldsymbol{h}(S \setminus F) \in \mathscr{F}_p$ because it is the union of two sets which are in $\mathscr{F}_p$. ■

In particular, this proves the following theorem from a different point of view to that done before, using $\boldsymbol{x} \to A\boldsymbol{x}$ being differentiable rather than $\boldsymbol{x} \to A\boldsymbol{x}$ being Lipschitz. Later on, is a theorem which says that Lipschitz implies differentiable a.e. However, it is also good to note that if $\boldsymbol{h}$ has a derivative on an open set U, it does not follow that $\boldsymbol{h}$ is Lipschitz.

Theorem 11.14.3 *Let A be a $p \times p$ matrix. Then if E is a Lebesgue measurable set, it follows that $A(E)$ is also a Lebesgue measurable set.*

I will also use the following fundamental assertion, Sard's lemma.

Lemma 11.14.4 *(Sard) Let U be an open set in $\mathbb{R}^p$. Let $\boldsymbol{h} : U \to \mathbb{R}^p$ be continuous and let $\boldsymbol{h}$ be differentiable on $A \subseteq U$. Let $Z \equiv \{\boldsymbol{x} \in A : \det D\boldsymbol{h}(\boldsymbol{x}) = 0\}$. Then $m_p(\boldsymbol{h}(Z)) = 0$.*

Proof: The balls in the following argument come from $\|\cdot\|_{\infty}$.

Suppose first that A is bounded. Let $\varepsilon > 0$ be given. Also let $V \supseteq Z$ with $V \subseteq U$ open, and $m_p(Z) + \varepsilon > m_p(V)$. Now let $\boldsymbol{x} \in Z$. Then since $\boldsymbol{h}$ is differentiable at $\boldsymbol{x}$, there exists $\delta_{\boldsymbol{x}} > 0$ such that if $r < \delta_{\boldsymbol{x}}$, then $B(\boldsymbol{x}, r) \subseteq V$ and also,

$$\boldsymbol{h}(B(\boldsymbol{x}, r)) \subseteq \boldsymbol{h}(\boldsymbol{x}) + D\boldsymbol{h}(\boldsymbol{x})(B(\boldsymbol{0}, r)) + B(\boldsymbol{0}, r\eta), \ \eta < 1.$$

Regard $D\boldsymbol{h}(\boldsymbol{x})$ as an $n \times n$ matrix, the matrix of the linear transformation $D\boldsymbol{h}(\boldsymbol{x})$ with respect to the usual coordinates. Since $\boldsymbol{x} \in Z$, it follows that there exists an invertible matrix M such that $MD\boldsymbol{h}(\boldsymbol{x})$ is in row reduced echelon form with a row of zeros on the bottom. Therefore,

$$m_p(M(\boldsymbol{h}(B(\boldsymbol{x}, r)))) \leq m_p(MD\boldsymbol{h}(\boldsymbol{x})(B(\boldsymbol{0}, r)) + MB(\boldsymbol{0}, r\eta)) \tag{11.21}$$

The diameter of $M\boldsymbol{Dh}(\boldsymbol{x})(B(\mathbf{0},r))$ is no larger than $\|M\|\|\boldsymbol{Dh}(\boldsymbol{x})\|2r$ and it lies in $\mathbb{R}^{p-1}\times\{0\}$. The diameter of $M\boldsymbol{B}(\mathbf{0},r\eta)$ is no more than $\|M\|(2r\eta)$. Enclose $M(\boldsymbol{h}(B(\boldsymbol{x},r)))$ in a rectangular p dimensional box having small height, sides parallel to the coordinate axes, the measure of the right side in 11.21 is no more than

$$[(\|M\|\|\boldsymbol{Dh}(\boldsymbol{x})\|2r+\|M\|(2\eta))r]^{p-1}(r\eta)\leq C(\|M\|,\|\boldsymbol{Dh}(\boldsymbol{x})\|)(2r)^p\eta$$

Hence from the change of variables formula for linear maps,

$$|\det(M)|\,m_p(\boldsymbol{h}(B(\boldsymbol{x},r)))=m_p(M\boldsymbol{h}(B(\boldsymbol{x},r)))\leq C(\|M\|,\|\boldsymbol{Dh}(\boldsymbol{x})\|)(2r)^p\eta$$

and so it follows that

$$m_p(\boldsymbol{h}(B(\boldsymbol{x},r)))\leq\frac{C(\|M\|,\|\boldsymbol{Dh}(\boldsymbol{x})\|)(2r)^p\eta}{|\det(M)|}=\frac{C(\|M\|,\|\boldsymbol{Dh}(\boldsymbol{x})\|)\eta}{|\det(M)|}(2r)^p$$

Then letting $\delta_{\boldsymbol{x}}$ be still smaller if necessary, corresponding to sufficiently small η ,

$$m_p(\boldsymbol{h}(B(\boldsymbol{x},r)))\leq\varepsilon m_p(B(\boldsymbol{x},r)).$$

The balls of this form constitute a Vitali cover of Z. Hence, by the covering theorem Corollary 9.12.5, there exists $\{B_i\}_{i=1}^{\infty}$, $B_i=B_i(\boldsymbol{x}_i,r_i)$, a collection of disjoint balls, each of which is contained in V, such that $m_p(\boldsymbol{h}(B_i))\leq\varepsilon m_p(B_i)$ and $m_p(Z\setminus\cup_iB_i)=0$. Hence from Lemma 11.14.1,

$$m_p(\boldsymbol{h}(Z)\setminus\cup_i\boldsymbol{h}(B_i))\leq m_p(\boldsymbol{h}(Z\setminus\cup_iB_i))=0$$

Therefore,

$$m_p(\boldsymbol{h}(Z))\leq\sum_i m_p(\boldsymbol{h}(B_i))\leq\varepsilon\sum_i m_p(B_i)\leq\varepsilon(m_p(V))\leq\varepsilon(m_p(Z)+\varepsilon).$$

Since ε is arbitrary, this shows $m_p(\boldsymbol{h}(Z))=0$. What if A is not bounded? Then consider $Z_n=Z\cap B(\mathbf{0},n)\subseteq A\cap B(\mathbf{0},n)$. From what was just shown, $\boldsymbol{h}(Z_n)$ has measure 0 and so it follows that $\boldsymbol{h}(Z)$ also does, being the countable union of sets of measure zero. ■

11.15 Change of Variables, Nonlinear Maps

Now with this preparation, we can prove a good change of variables formula. First is a lemma which is likely familiar by now.

Lemma 11.15.1 *Let $\boldsymbol{h}:\Omega\to\mathbb{R}^p$ where $(\Omega,\mathscr{F})$ is a measurable space and suppose $\boldsymbol{h}$ is continuous. Then $\boldsymbol{h}^{-1}(B)\in\mathscr{F}$ whenenver B is a Borel set.*

Proof: Measurability applied to components of $\boldsymbol{h}$ shows that $\boldsymbol{h}^{-1}(U)\in\mathscr{F}$ whenever U is an open set. If $\mathscr{G}$ is consists of the subsets G of $\mathbb{R}^p$ for which $\boldsymbol{h}^{-1}(G)\in\mathscr{F}$, then $\mathscr{G}$ is a σ algebra and $\mathscr{G}$ contains the open sets. Hence it also contains the Borel sets. ■

Definition 11.15.2 *Let $\boldsymbol{h}:U\to\boldsymbol{h}(U)$ be continuous and let $H\subseteq U$ be measurable with $m_p(\boldsymbol{h}(U\setminus H))=0$ and that $\boldsymbol{h}$ is one to one and differentiable on H. Define $\lambda(E)\equiv m_p(\boldsymbol{h}(E\cap H))$.*

Lemma 11.15.3 *λ is a well defined measure and $\lambda \ll m_p$. Also*

$$\lambda(E) = m_p(\boldsymbol{h}(E \cap H)) = m_p(\boldsymbol{h}(E))$$

for E a measurable subset of U.

$$\lambda(E) = \int_E D_{m_p}\lambda(\boldsymbol{x})\, dm_p(\boldsymbol{x}) \tag{11.22}$$

For $\boldsymbol{x} \notin H, D_{m_p}\lambda(\boldsymbol{x}) = 0$ a.e. Also $\lambda(E) = \lambda\left(E \cap Z^C\right)$. $Z \equiv \{\boldsymbol{x} \in H : \det(D\boldsymbol{h}) = \mathbf{0}\}$.

Proof: Note that $\boldsymbol{h}(E) = \boldsymbol{h}(E \cap H) \cup \boldsymbol{h}(E \setminus H)$ and by assumption, the set on the right has measure zero so $\boldsymbol{h}(E)$ is also measurable for any measurable $E \subseteq U$. $\boldsymbol{h}(E \cap H)$ is m_p measurable by Lemma 11.14.2. Thus $m_p(\boldsymbol{h}(E)) = m_p(\boldsymbol{h}(E \cap H))$.

Say the E_i are disjoint. $\lambda(\cup_i E_i) \equiv m_p(\boldsymbol{h}(\cup_i E_i \cap H)) = \sum_i m_p(\boldsymbol{h}(E_i \cap H)) = \sum_i \lambda(E_i)$. If $m_p(E) = 0$, then $\lambda(E) \equiv m_p(\boldsymbol{h}(E \cap H)) = 0$ because of Lemma 11.14.1. 11.22 follows from Proposition 11.6.2. $\lambda(U \setminus H) \equiv m_p(\boldsymbol{h}((U \setminus H) \cap H)) = 0$. Thus $\lambda(U \setminus H) = \int_{U \setminus H} D_{m_p}\lambda(\boldsymbol{x})\, dm_p(\boldsymbol{x}) = 0$ so indeed it follows that $D_{m_p}\lambda(\boldsymbol{x}) = 0$ off H.

As to the last claim, $\lambda(E) \equiv m_p(\boldsymbol{h}(E \cap H))$. By Sard's theorem, Theorem 11.14.4,

$$\begin{aligned} m_p(\boldsymbol{h}(E \cap H)) &= m_p\left(\boldsymbol{h}\left(E \cap H \cap Z^C\right) \cup \boldsymbol{h}(E \cap H \cap Z)\right) \\ &\leq m_p\left(\boldsymbol{h}\left(E \cap H \cap Z^C\right)\right) + m_p(\boldsymbol{h}(Z)) \\ &= m_p\left(\boldsymbol{h}\left(E \cap H \cap Z^C\right)\right) \leq m_p(\boldsymbol{h}(E \cap H)) \end{aligned}$$

Thus $\lambda(E) = m_p\left(\boldsymbol{h}\left(E \cap H \cap Z^C\right)\right) = \lambda\left(E \cap Z^C\right)$ as claimed. ■

Recall Lemma 8.10.1 which depended on the Brouwer fixed point theorem.

Lemma 11.15.4 *Let $\boldsymbol{f}$ be continuous and map $\overline{B(\boldsymbol{p},r)} \subseteq \mathbb{R}^p$ to $\mathbb{R}^p$. Suppose that for all $\boldsymbol{x} \in \overline{B(\boldsymbol{p},r)}, |\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{x}| < \varepsilon r$. Then it follows that $\boldsymbol{f}\left(\overline{B(\boldsymbol{p},r)}\right) \supseteq B(\boldsymbol{p},(1-\varepsilon)r)$.*

Lemma 11.15.5 *From Definition 11.15.2, if $\det D\boldsymbol{h}(\boldsymbol{x}) \neq 0, D_{m_p}\lambda(\boldsymbol{x}) = |\det D\boldsymbol{h}(\boldsymbol{x})|$ for any $\boldsymbol{x} \in H$.*

Proof: From the definition of differentiability at $\boldsymbol{x}$, for all r small enough, and also that the balls are in U,

$$\begin{aligned} \frac{\lambda(B(\boldsymbol{x},r))}{m_p(B(\boldsymbol{x},r))} &= \frac{m_p(\boldsymbol{h}(B(\boldsymbol{x},r) \cap H))}{m_p(B(\boldsymbol{x},r))} = \frac{m_p(\boldsymbol{h}(B(\boldsymbol{x},r)))}{m_p(B(\boldsymbol{x},r))} \\ &\leq \frac{m_p\left(\boldsymbol{h}(\boldsymbol{x}) + D\boldsymbol{h}(\boldsymbol{x})B(\mathbf{0},r) + D\boldsymbol{h}(\boldsymbol{x})D\boldsymbol{h}(\boldsymbol{x})^{-1}B(\mathbf{0},\delta r)\right)}{m_p(B(\boldsymbol{x},r))} \\ &= \frac{m_p(D\boldsymbol{h}(\boldsymbol{x})B(\mathbf{0},r) + D\boldsymbol{h}(\boldsymbol{x})B(\mathbf{0},\varepsilon r))}{m_p(B(\boldsymbol{x},r))} \\ &\leq \frac{m_p(D\boldsymbol{h}(\boldsymbol{x})(B(\mathbf{0},r(1+\varepsilon))))}{m_p(B(\boldsymbol{x},r))} = |\det(D\boldsymbol{h}(\boldsymbol{x}))|(1+\varepsilon)^p \end{aligned}$$

if δ is small enough. Thus, since ε is arbitrary, $D_{m_p}\lambda(\boldsymbol{x}) \leq |\det(D\boldsymbol{h}(\boldsymbol{x}))|$. It remains to consider the other direction. This very interesting case follows Rudin [49].

If $\|\boldsymbol{v}\| \leq r$ for suitably small r that the ball $\overline{B(\boldsymbol{x},r)} \subseteq U$,

$$\boldsymbol{h}(\boldsymbol{x}+\boldsymbol{v}) = \boldsymbol{h}(\boldsymbol{x}) + D\boldsymbol{h}(\boldsymbol{x})\boldsymbol{v} + D\boldsymbol{h}(\boldsymbol{x})D\boldsymbol{h}(\boldsymbol{x})^{-1}\boldsymbol{o}(\boldsymbol{v})$$

and so $\left\|D\boldsymbol{h}(\boldsymbol{x})^{-1}(\boldsymbol{h}(\boldsymbol{x}+\boldsymbol{v})-\boldsymbol{h}(\boldsymbol{x}))-\boldsymbol{v}\right\| = \left\|D\boldsymbol{h}(\boldsymbol{x})^{-1}\boldsymbol{o}(\boldsymbol{v})\right\| = \|\boldsymbol{o}(\boldsymbol{v})\|$. Thus from Lemma 8.10.1, if r is chosen sufficiently small,

$$\left\|D\boldsymbol{h}(\boldsymbol{x})^{-1}(\boldsymbol{h}(\boldsymbol{x}+\boldsymbol{v})-\boldsymbol{h}(\boldsymbol{x}))-\boldsymbol{v}\right\| < \varepsilon r$$

and so, from the Lemma 11.15.4, $B(\mathbf{0},(1-\varepsilon)r) \subseteq D\boldsymbol{h}(\boldsymbol{x})^{-1}\left(\boldsymbol{h}\left(\boldsymbol{x}+\overline{B(\mathbf{0},r)}\right)-\boldsymbol{h}(\boldsymbol{x})\right)$. Therefore, since $\boldsymbol{h}(B(\boldsymbol{x},r)) = \boldsymbol{h}\left(\overline{B(\boldsymbol{x},r)}\right)$,

$$m_p(D\boldsymbol{h}(\boldsymbol{x})B(\mathbf{0},(1-\varepsilon)r)) \leq m_p(\boldsymbol{h}(B(\boldsymbol{x},r)))$$

and so, from the above, at every $\boldsymbol{x}$ where $D\boldsymbol{h}(\boldsymbol{x})^{-1}$ exists,

$$D_{m_p}(\lambda)(\boldsymbol{x}) \in \left[\frac{m_p(D\boldsymbol{h}(\boldsymbol{x})B(\boldsymbol{x},(1-\varepsilon)r))}{m_p(B(\boldsymbol{x},r))}, \frac{m_p(D\boldsymbol{h}(\boldsymbol{x})B(\boldsymbol{x},(1+\varepsilon)r))}{m_p(B(\boldsymbol{x},r))}\right]$$

It follows from this that since ε is arbitrary, $D_{m_p}(\lambda)(\boldsymbol{x}) = |\det(D\boldsymbol{h}(\boldsymbol{x}))|$ at all points $\boldsymbol{x}$ where $|\det(D\boldsymbol{h}(\boldsymbol{x}))| \neq 0$. ■

Now here is the change of variables formula.

Theorem 11.15.6 *Let U be an open set and let $\boldsymbol{h}: U \to \boldsymbol{h}(U)$ be continuous and differentiable on the measurable $H \subseteq U$ such that $\boldsymbol{h}(U \setminus H)$ has measure zero. Also let $\boldsymbol{h}$ be one to one on H. Then if $f \geq 0$ is Lebesgue measurable,*

$$\int_{\boldsymbol{h}(H)} f(y)\,dm_p = \int_H f(\boldsymbol{h}(\boldsymbol{x}))\,|\det(D\boldsymbol{h}(\boldsymbol{x}))|\,dm_p$$

Proof: Letting Z be those $\boldsymbol{x} \in H$ for which $|\det(D\boldsymbol{h}(\boldsymbol{x}))| = 0$, it follows from Sard's lemma and Lemmas 11.15.5 and 11.15.3, the Radon Nikodym theorem and the characterization as the symmetric derivative, if $E \subseteq U$,

$$\begin{aligned}\lambda(E) &= \lambda\left(E \cap Z^C\right) = \int_{E\cap Z^C} D_{m_p}\lambda(\boldsymbol{x})\,dm_p \\ &= \int_{E\cap Z^C} |\det(D\boldsymbol{h}(\boldsymbol{x}))|\,dm_p(\boldsymbol{x}) = \int_E |\det(D\boldsymbol{h}(\boldsymbol{x}))|\,dm_p(\boldsymbol{x})\end{aligned}$$

Now let F be a Borel set so $\boldsymbol{h}^{-1}(F) \equiv \{\boldsymbol{x} \in H : \boldsymbol{h}(\boldsymbol{x}) \in F\}$ is measurable. Then the above shows that

$$\begin{aligned}&\int_H \mathcal{X}_F(\boldsymbol{h}(\boldsymbol{x}))\,|\det(D\boldsymbol{h}(\boldsymbol{x}))|\,dm_p(\boldsymbol{x}) = \int_{\boldsymbol{h}^{-1}(F)} |\det(D\boldsymbol{h}(\boldsymbol{x}))|\,dm_p(\boldsymbol{x}) \\ = \;& \lambda\left(\boldsymbol{h}^{-1}(F)\right) \equiv m_p(F) = m_p\left(\boldsymbol{h}\left(\boldsymbol{h}^{-1}(F)\right)\right) = \int_{\boldsymbol{h}(H)} \mathcal{X}_F(\boldsymbol{y})\,dm_p(\boldsymbol{y})\end{aligned}$$

Thus, approximating a nonnegative Borel measurable function g with simple functions and using the monotone convergence theorem and the above,

$$\int_{\boldsymbol{h}(H)} g(\boldsymbol{y})\,dm_p = \int_H g(\boldsymbol{h}(\boldsymbol{x}))\,|\det(D\boldsymbol{h}(\boldsymbol{x}))|\,dm_p(\boldsymbol{x})$$

It only remains to verify this is true for $g \to f$ where f is Lebesgue measurable. From regularity, there exist nonnegative Borel measurable functions k, g such that $k(\boldsymbol{y}) \leq f(\boldsymbol{y}) \leq g(\boldsymbol{y})$ with equality holding off a set of m_p measure zero. Then

$$\int_{\boldsymbol{h}(H)} f(\boldsymbol{y})\, dm_p = \int_{\boldsymbol{h}(H)} k(\boldsymbol{y})\, dm_p = \int_H k(\boldsymbol{h}(\boldsymbol{x}))\, |\det(D\boldsymbol{h}(\boldsymbol{x}))|\, dm_p(\boldsymbol{x})$$

$$\leq \int_H g(\boldsymbol{h}(\boldsymbol{x}))\, |\det(D\boldsymbol{h}(\boldsymbol{x}))|\, dm_p(\boldsymbol{x}) = \int_{\boldsymbol{h}(H)} g(\boldsymbol{y})\, dm_p = \int_{\boldsymbol{h}(H)} f(\boldsymbol{y})\, dm_p \tag{11.23}$$

Therefore,

$$k(\boldsymbol{h}(\boldsymbol{x}))\, |\det(D\boldsymbol{h}(\boldsymbol{x}))| \leq f(\boldsymbol{h}(\boldsymbol{x}))\, |\det(D\boldsymbol{h}(\boldsymbol{x}))| \leq g(\boldsymbol{h}(\boldsymbol{x}))\, |\det(D\boldsymbol{h}(\boldsymbol{x}))|$$

and all functions are equal off a set of measure zero because of 11.23, the two on the ends being Borel measurable functions of $\boldsymbol{x}$. By completeness of Lebesgue measure, $\boldsymbol{x} \to f(\boldsymbol{h}(\boldsymbol{x}))\, |\det(D\boldsymbol{h}(\boldsymbol{x}))|$ is Lebesgue measurable. Note this does not say that $\boldsymbol{x} \to f(\boldsymbol{h}(\boldsymbol{x}))$ is. Then from 11.23, it follows that $\int_U f(\boldsymbol{h}(\boldsymbol{x}))\, |\det(D\boldsymbol{h}(\boldsymbol{x}))|\, dm_p(\boldsymbol{x})$ equals $\int_V f(\boldsymbol{y})\, dm_p$. ■

Rademacher's theorem, presented later, states that Lipschitz functions are differentiable a.e. and Theorem 11.13.1 says that Lipschitz maps take sets of measure zero to sets of measure zero. Thus, the above corollary implies the following corollary, a case of the area formula presented later.

Corollary 11.15.7 *Suppose $\boldsymbol{h}$ is one to one and Lipschitz on U. Then let H be those points where $\boldsymbol{h}$ is differentiable, a.e. point of U. Then if f is Lebesgue measurable and nonnegative,*

$$\begin{aligned} \int_{\boldsymbol{h}(U)} f\, dm_p &= \int_{\boldsymbol{h}(H)} f\, dm_p = \int_H f(\boldsymbol{h}(\boldsymbol{x}))\, |\det D\boldsymbol{h}(\boldsymbol{x})|\, dm_p \\ &= \int_U f(\boldsymbol{h}(\boldsymbol{x}))\, |\det D\boldsymbol{h}(\boldsymbol{x})|\, dm_p \end{aligned}$$

11.16 Mappings which are not One to One

Now suppose $\boldsymbol{h}$ is $C^1(U)$, not necessarily one to one. For

$$U_+ \equiv \{\boldsymbol{x} \in U : |\det D\boldsymbol{h}(\boldsymbol{x})| > 0\}$$

and Z the set where $|\det D\boldsymbol{h}(\boldsymbol{x})| = 0$, Lemma 11.14.4 implies $m_p(\boldsymbol{h}(Z)) = 0$. For $\boldsymbol{x} \in U_+$, the inverse function theorem implies there exists an open set $B_{\boldsymbol{x}} \subseteq U_+$, such that $\boldsymbol{h}$ is one to one on $B_{\boldsymbol{x}}$.

Let $\{B_i\}$ be a countable subset of $\{B_{\boldsymbol{x}}\}_{\boldsymbol{x} \in U_+}$ such that $U_+ = \cup_{i=1}^{\infty} B_i$. Let $E_1 = B_1$. If $E_1, \cdots, E_k$ have been chosen, $E_{k+1} = B_{k+1} \setminus \cup_{i=1}^{k} E_i$. Thus

$$\cup_{i=1}^{\infty} E_i = U_+, \ \ \boldsymbol{h} \text{ is one to one on } E_i, \ \ E_i \cap E_j = \emptyset,$$

and each E_i is a Borel set contained in the open set B_i. Now define

$$n(\boldsymbol{y}) \equiv \sum_{i=1}^{\infty} \mathcal{X}_{\boldsymbol{h}(E_i)}(\boldsymbol{y}) + \mathcal{X}_{\boldsymbol{h}(Z)}(\boldsymbol{y}).$$

The set $\boldsymbol{h}(E_i), \boldsymbol{h}(Z)$ are measurable by Lemma 11.14.2. Thus $n(\cdot)$ is measurable.

Lemma 11.16.1 *Let $F \subseteq \boldsymbol{h}(U)$ be measurable. Then*

$$\int_{\boldsymbol{h}(U)} n(\boldsymbol{y})\mathcal{X}_F(\boldsymbol{y})dm_p = \int_U \mathcal{X}_F(\boldsymbol{h}(\boldsymbol{x}))|\det D\boldsymbol{h}(\boldsymbol{x})|dm_p.$$

Proof: Using Lemma 11.14.4 and the Monotone Convergence Theorem

$$\int_{\boldsymbol{h}(U)} n(\boldsymbol{y})\mathcal{X}_F(\boldsymbol{y})dm_p = \int_{\boldsymbol{h}(U)} \left(\sum_{i=1}^{\infty} \mathcal{X}_{\boldsymbol{h}(E_i)}(\boldsymbol{y}) + \overbrace{\mathcal{X}_{\boldsymbol{h}(Z)}(\boldsymbol{y})}^{m_p(\boldsymbol{h}(Z))=0} \right) \mathcal{X}_F(\boldsymbol{y})dm_p$$

$$= \sum_{i=1}^{\infty} \int_{\boldsymbol{h}(U)} \mathcal{X}_{\boldsymbol{h}(E_i)}(\boldsymbol{y})\mathcal{X}_F(\boldsymbol{y})dm_p = \sum_{i=1}^{\infty} \int_{\boldsymbol{h}(B_i)} \mathcal{X}_{\boldsymbol{h}(E_i)}(\boldsymbol{y})\mathcal{X}_F(\boldsymbol{y})dm_p =$$

$$\sum_{i=1}^{\infty} \int_{B_i} \mathcal{X}_{E_i}(\boldsymbol{x})\mathcal{X}_F(\boldsymbol{h}(\boldsymbol{x}))|\det D\boldsymbol{h}(\boldsymbol{x})|dm_p = \sum_{i=1}^{\infty} \int_U \mathcal{X}_{E_i}(\boldsymbol{x})\mathcal{X}_F(\boldsymbol{h}(\boldsymbol{x}))|\det D\boldsymbol{h}(\boldsymbol{x})|dm_p$$

$$= \int_U \sum_{i=1}^{\infty} \mathcal{X}_{E_i}(\boldsymbol{x})\mathcal{X}_F(\boldsymbol{h}(\boldsymbol{x}))|\det D\boldsymbol{h}(\boldsymbol{x})|dm_p$$

$$= \int_{U_+} \mathcal{X}_F(\boldsymbol{h}(\boldsymbol{x}))|\det D\boldsymbol{h}(\boldsymbol{x})|dm_p = \int_U \mathcal{X}_F(\boldsymbol{h}(\boldsymbol{x}))|\det D\boldsymbol{h}(\boldsymbol{x})|dm_p. \blacksquare$$

Definition 11.16.2 *For $\boldsymbol{y} \in \boldsymbol{h}(U)$, define a function, #, according to the formula* $\#(\boldsymbol{y}) \equiv$ *number of elements in* $\boldsymbol{h}^{-1}(\boldsymbol{y})$.

Observe that $\#(\boldsymbol{y}) = n(\boldsymbol{y})$ a.e. because $n(\boldsymbol{y}) = \#(\boldsymbol{y})$ if $\boldsymbol{y} \notin \boldsymbol{h}(Z)$, a set of measure 0. Therefore, # is a measurable function because of completeness of Lebesgue measure.

Theorem 11.16.3 *Let $g \geq 0$, g measurable, and let $\boldsymbol{h}$ be $C^1(U)$. Then*

$$\int_{\boldsymbol{h}(U)} \#(\boldsymbol{y})g(\boldsymbol{y})dm_p = \int_U g(\boldsymbol{h}(\boldsymbol{x}))|\det D\boldsymbol{h}(\boldsymbol{x})|dm_p. \tag{11.24}$$

Proof: From $\#(\boldsymbol{y}) = n(\boldsymbol{y})$ and Lemma 11.16.1, 11.24 holds for all g, a nonnegative simple function. Approximating an arbitrary measurable nonnegative function, g, with an increasing pointwise convergent sequence of simple functions and using the monotone convergence theorem, yields 11.24 for an arbitrary nonnegative measurable function, g. $\blacksquare$

Note how $\#(\boldsymbol{y}) \geq 1$ on $\boldsymbol{h}(Z)$ where Z is as above, the set where $|\det D\boldsymbol{h}(\boldsymbol{x})| = 0$ so the above formula could be used to conclude that $m_p(\boldsymbol{h}(Z)) = 0$.

11.17 Spherical Coordinates

As usual, S^{p-1} is the unit sphere, the boundary of the unit ball $B(\mathbf{0},1)$ in $\mathbb{R}^p$. It is a metric space with respect to the usual notion of distance which it inherits from being a part of $\mathbb{R}^p$. Then $(0,\infty) \times S^{p-1}$ is also a metric space with the metric

$$d((\rho,\boldsymbol{\omega}),(\hat{\rho},\hat{\boldsymbol{\omega}})) \equiv \max\{|\rho-\hat{\rho}|, |\boldsymbol{\omega}-\hat{\boldsymbol{\omega}}|\}$$

Indeed, this kind of thing delivers a metric for an arbitrary finite product of metric spaces. See Problem 6 on Page 93.

Definition 11.17.1 *Let $\lambda : \mathbb{R}^p \setminus \{\mathbf{0}\} \to (0,\infty) \times S^{p-1}$ be defined by*

$$\lambda(\boldsymbol{x}) \equiv \left(|\boldsymbol{x}|, \frac{\boldsymbol{x}}{|\boldsymbol{x}|} \right)$$

Then with this definition, the following is true.

Lemma 11.17.2 *Let λ be as defined above. Then λ is one to one, onto, and continuous with continuous inverse.*

Proof: First of all, it is obviously onto. Indeed, if $(\rho,\boldsymbol{\omega}) \in (0,\infty) \times S^{p-1}$, consider $\boldsymbol{x} \equiv \rho\boldsymbol{\omega}$. Why is this one to one? If $\boldsymbol{x} \neq \hat{\boldsymbol{x}}$, then there are two cases. It might be that $|\boldsymbol{x}| \neq |\hat{\boldsymbol{x}}|$ and in this case, clearly $\lambda(\boldsymbol{x}) \neq \lambda(\hat{\boldsymbol{x}})$. The other case is that $|\boldsymbol{x}| = |\hat{\boldsymbol{x}}| = \rho$ but these two vectors $\boldsymbol{x}, \hat{\boldsymbol{x}}$ are not equal. In this case, $\frac{\boldsymbol{x}}{|\boldsymbol{x}|} - \frac{\hat{\boldsymbol{x}}}{|\hat{\boldsymbol{x}}|} = \frac{1}{\rho}(\boldsymbol{x} - \hat{\boldsymbol{x}}) \neq \mathbf{0}$. Thus λ is one to one.

Is λ continuous? Suppose $\boldsymbol{x}_n \to \boldsymbol{x} \neq \mathbf{0}$. Does $\lambda(\boldsymbol{x}_n) \to \lambda(\boldsymbol{x})$? First of all, the triangle inequality shows that $|\boldsymbol{x}_n| \to |\boldsymbol{x}|$. It only remains to verify $\frac{\boldsymbol{x}_n}{|\boldsymbol{x}_n|} \to \frac{\boldsymbol{x}}{|\boldsymbol{x}|}$. This is clearly the case because

$$\left| \frac{\boldsymbol{x}_n}{|\boldsymbol{x}_n|} - \frac{\boldsymbol{x}}{|\boldsymbol{x}|} \right| = \left| \frac{|\boldsymbol{x}|\,\boldsymbol{x}_n - |\boldsymbol{x}_n|\,\boldsymbol{x}}{|\boldsymbol{x}_n|\,|\boldsymbol{x}|} \right| \to \left| \frac{|\boldsymbol{x}|\,\boldsymbol{x} - |\boldsymbol{x}|\,\boldsymbol{x}}{|\boldsymbol{x}|\,|\boldsymbol{x}|} \right| = 0$$

Is λ^{-1} also continuous? One could show this directly or observe that λ^{-1} is automatically continuous on $\left[\frac{1}{n}, n\right] \times S^{p-1}$ because this is a compact set. Indeed, $\left[\frac{1}{n}, n\right] \times S^{p-1} = \lambda\left(\left\{\boldsymbol{x} \in \mathbb{R}^p : \frac{1}{n} \leq |\boldsymbol{x}| \leq n\right\}\right)$. If $\lambda(\boldsymbol{x}_n) \to \lambda(\boldsymbol{x})$, does it follow that $\boldsymbol{x}_n \to \boldsymbol{x}$? If not, there exists a subsequence, still denoted as $\boldsymbol{x}_n$ such that $\boldsymbol{x}_n \to \boldsymbol{y} \neq \boldsymbol{x}$. But then, by continuity of λ, $\lambda(\boldsymbol{x}_n) \to \lambda(\boldsymbol{y})$ and so $\lambda(\boldsymbol{y}) = \lambda(\boldsymbol{x})$ which does not occur because λ is one to one.

It follows, since $(0,\infty) \times S^{p-1} = \cup_n \left[\frac{1}{n}, n\right] \times S^{p-1}$, that λ^{-1} is continuous. ■

Thus the open sets for $(0,\infty) \times S^{p-1}$ are all of the form $\lambda(U)$ where U is open in $\mathbb{R}^p \setminus \{\mathbf{0}\}$. Also, the open sets of $\mathbb{R}^p \setminus \{\mathbf{0}\}$ are of the form $\lambda^{-1}(V)$ where V is an open set of $(0,\infty) \times S^{p-1}$. One can replace the word "open" with the word "Borel" in the previous observation.

Motivated by familiar formulas for the area of a sphere and the circumference of a circle, here is a definition of a surface measure defined on the Borel sets of S^{p-1}.

Definition 11.17.3 *Let E be a Borel set on S^{p-1}. Then*

$$\lambda^{-1}((0,1] \times E) \equiv \{\rho\boldsymbol{\omega} : \rho \in (0,1], \boldsymbol{\omega} \in E\}$$

will be a part of the unit ball formed from the cone starting at $\mathbf{0}$ and extending to the points of E, leaving out $\mathbf{0}$. Since $(0,1] \times E$ is a Borel set in $(0,\infty) \times S^{p-1}$ thanks to Problem 4 on Page 254, this cone just described is a Borel set in $\mathbb{R}^p$. Then

$$\sigma(E) \equiv pm_p\left(\lambda^{-1}((0,1] \times E)\right)$$

This is obviously a measure on the Borel sets of S^{p-1}.

Is this even a good idea? Note $m_p\left(\lambda^{-1}(\{r\} \times E)\right) = 0$ because $\lambda^{-1}(\{r\} \times E)$ is just a part of the sphere of radius r which has m_p measure zero. The reason this is so is as follows. Letting $\alpha_p \equiv m_p(B(\mathbf{0},1))$, the sphere of radius r is contained in $B(\mathbf{0}, r+\varepsilon) \setminus B(\mathbf{0}, r-\varepsilon)$ and so the sphere has m_p measure no more than $\alpha_p((r+\varepsilon)^p - (r-\varepsilon)^p)$ for every $\varepsilon > 0$.

Lemma 11.17.4 *Let G be a Borel set in $(0,\infty)\times S^{p-1}$. Then*

$$m_p\left(\lambda^{-1}(G)\right)=\int_0^\infty\int_{S^{p-1}}\mathcal{X}_G(\rho,\omega)\rho^{p-1}d\sigma d\rho \tag{11.25}$$

and the iterated integrals make sense.

Proof: Let $\mathcal{K}\equiv\left\{I\times E : I \text{ is an interval in } (0,\infty) \text{ and } E \text{ is Borel in } S^{p-1}\right\}$. This is a π system and each set of $\mathcal{K}$ is a Borel set. Then if I is one of these intervals, having end points $a<b$,

$$\int_0^\infty\int_{S^{p-1}}\mathcal{X}_{I\times E}(\rho,\omega)\rho^{p-1}d\sigma d\rho=\int_a^b\rho^{p-1}\int_E d\sigma d\rho=\sigma(E)\left(\frac{b^p}{p}-\frac{a^p}{p}\right)$$

$$\begin{aligned}&= p\,m_p\left(\lambda^{-1}((0,1]\times E)\right)\left(\frac{b^p}{p}-\frac{a^p}{p}\right)=m_p\left(\lambda^{-1}((0,1]\times E)\right)(b^p-a^p)\\&= m_p\left(\lambda^{-1}((a,b)\times E)\right)=m_p\left(\lambda^{-1}(I\times E)\right)\end{aligned}$$

Let $\mathcal{G}$ denote those Borel sets G in $(0,\infty)\times S^{p-1}$ for which, $G_n\equiv G\cap(0,n)\times S^{p-1}$,

$$m_p\left(\lambda^{-1}(G_n)\right)=\int_0^\infty\int_{S^{p-1}}\mathcal{X}_{G_n}(\rho,\omega)\rho^{p-1}d\sigma d\rho$$

the iterated integrals making sense. It is routine to verify that $\mathcal{G}$ is closed with respect to complements and countable disjoint unions. It was also shown above that it contains $\mathcal{K}$. By Dynkin's lemma, Lemma 9.3.2, $\mathcal{G}$ equals the Borel sets in $(0,\infty)\times S^{p-1}$. Now use the monotone convergence theorem. ■

Theorem 11.17.5 *Let f be a Borel measurable nonnegative function. Then*

$$\int f dm_p=\int_0^\infty\int_{S^{p-1}}f(\rho\omega)\rho^{p-1}d\sigma d\rho \tag{11.26}$$

Proof: From the above lemma, if F is an arbitrary Borel set, it has the same measure as $F\cap(\mathbb{R}^p\setminus\{\mathbf{0}\})$ so there is no loss of generality in assuming $\mathbf{0}\notin F$.

$$\int_{\mathbb{R}^p}\mathcal{X}_F dm_p=m_p(F)=m_p\left(\lambda^{-1}(\lambda(F))\right)=\int_0^\infty\int_{S^{p-1}}\mathcal{X}_{\lambda(F)}(\rho,\omega)\rho^{p-1}d\sigma d\rho$$

$$=\int_0^\infty\int_{S^{p-1}}\mathcal{X}_F\left(\lambda^{-1}(\rho,\omega)\right)\rho^{p-1}d\sigma d\rho=\int_0^\infty\int_{S^{p-1}}\mathcal{X}_F(\rho\omega)\rho^{p-1}d\sigma d\rho$$

Now if f is nonnegative and Borel measurable, one can approximate using Borel simple functions increasing pointwise to f and use the monotone convergence theorem to obtain 11.26. ■

Note that by Theorem 10.14.8, you can interchange the order of integration in 11.25 if desired.

Example 11.17.6 *For what values of s is the integral*

$$\int_{B(\mathbf{0},R)}\left(1+|\boldsymbol{x}|^2\right)^s dy$$

bounded independent of R? Here $B(\mathbf{0},R)$ is the ball, $\{\boldsymbol{x}\in\mathbb{R}^p : |\boldsymbol{x}|\le R\}$.

I think you can see immediately that s must be negative but exactly how negative? It turns out it depends on p and using polar coordinates, you can find just exactly what is needed. From the polar coordinates formula above,

$$\begin{aligned}\int_{B(\mathbf{0},R)} \left(1+|\boldsymbol{x}|^2\right)^s dy &= \int_0^R \int_{S^{p-1}} \left(1+\rho^2\right)^s \rho^{p-1} d\sigma d\rho \\ &= C_p \int_0^R \left(1+\rho^2\right)^s \rho^{p-1} d\rho\end{aligned}$$

Now the very hard problem has been reduced to considering an easy one variable problem of finding when $\int_0^R \rho^{p-1}\left(1+\rho^2\right)^s d\rho$ is bounded independent of R. You need $2s+(p-1)<-1$ so you need $s<-p/2$.

11.18 Exercises

1. Show that $\int_0^\infty e^{-x^2}dx = \frac{1}{2}\sqrt{\pi}$. **Hint:** First verify the integral is finite. You might use monotone convergence theorem to do this. It is easier than the stuff you worried about in beginning calculus. Next let $I = \int_0^\infty e^{-x^2}dx$ so that $I^2 = \int_0^\infty \int_0^\infty e^{-\left(x^2+y^2\right)}dxdy$. Now change the variables using polar coordinates. It is all justified by the big change of variables theorem we have done. This becomes an easy problem when you do this.

2. Show that $\int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx = 1$. Here σ is a positive number called the standard deviation and μ is a number called the mean. **Hint:** Just use the above result to find $\int_{-\infty}^\infty e^{-x^2}dx$ and then change the variables in this one.

3. The Gamma function is $\Gamma(\alpha) \equiv \int_0^\infty e^{-t}t^{\alpha-1}dt$. Verify that this function is well defined in the sense that it is finite for all $\alpha > 0$. Next verify that $\Gamma(1) = 1 = \Gamma(2)$ and $\Gamma(x+1) = \Gamma(x)x$. Conclude that $\Gamma(n+1) = n!$ if n is an integer. Now consider $\Gamma(1/2)$. This is $\int_0^\infty e^{-t}t^{-1/2}dt$. Change the variables in this integral. You might let $t = u^2$. Then consider the above problem.

4. Let $p,q>0$ and define $B(p,q) = \int_0^1 x^{p-1}(1-x)^{q-1}$. Show that

$$\Gamma(p)\Gamma(q) = B(p,q)\Gamma(p+q).$$

   Hint: It is fairly routine if you start with the left side and proceed to change variables.

5. Let E be a Lebesgue measurable set in $\mathbb{R}$. Suppose $m(E) > 0$. Consider the set $E-E = \{x-y : x \in E, y \in E\}$. Show that $E-E$ contains an interval. This is an amazing result. Recall the case of the fat Cantor set which contained no intervals but had positive measure. **Hint:** Without loss of generality, you can assume E is bounded. Let $f(x) = \int \mathcal{X}_E(t)\mathcal{X}_E(x+t)dt$. Explain why f is continuous at 0 and $f(0) > 0$ and use continuity of translation in L^1. To see it is continuous,

$$\begin{aligned}|f(x)-f(\hat{x})| &\leq \int \mathcal{X}_E(t)\,|\mathcal{X}_E(x+t)-\mathcal{X}_E(\hat{x}+t)|\,dt \\ &\leq \int |\mathcal{X}_E(x+t)-\mathcal{X}_E(\hat{x}+t)|\,dt\end{aligned}$$

Now explain why this is small whenever $\hat{x}-x$ is small due to continuity of translation in $L^1(\mathbb{R})$. Thus $f(0)=m(E)>0$ and so by continuity, $f>0$ near 0. If the integral is nonzero, what can you say about the integrand? You must have for all $x\in(-\delta,\delta)$ both $x+t\in E$ and $t\in E$. Now consider this a little.

6. Does there exist a closed uncountable set which is contained in the set of irrational numbers? If so, explain why and if not, explain why. Thus this uncountable set has no rational number as a limit point.

7. Find the area of the bounded region R, determined by $5x+y=1, 5x+y=9, y=2x$, and $y=5x$.

8. Here are three vectors. $(1,2,3)^T, (1,0,1)^T$, and $(2,1,0)^T$. These vectors determine a parallelepiped, R, which is occupied by a solid having density $\rho=y$. Find the mass of this solid. To find the mass of the solid, you integrate the density. Thus, if P is this parallelepiped, the mass is $\int_P y dm_3$ **Hint:** Let $\boldsymbol{h}:[0,1]^3\to P$ be given by $\boldsymbol{h}(t_1,t_2,t_3)=t_1\begin{pmatrix}1\\2\\3\end{pmatrix}+t_2\begin{pmatrix}1\\0\\1\end{pmatrix}+t_3\begin{pmatrix}2\\1\\0\end{pmatrix}$ then by definition of what is meant by a parallelepiped, $\boldsymbol{h}\left([0,1]^2\right)=P$ and $\boldsymbol{h}$ is one to one and onto.

9. Find the volume of the ellipsoid $\frac{x^2}{a^2}+\frac{y^2}{b^2}+\frac{z^2}{c^2}=1$. **Hint:** Change variables to reduce to finding the volume of a sphere of radius 1 which you know is $(4/3)\pi$.

10. Suppose $f,g\in L^1(\mathbb{R}^p)$. Define $f*g(\boldsymbol{x})$ by $\int f(\boldsymbol{x}-\boldsymbol{y})g(\boldsymbol{y})dm_p(y)$. First show using the preceding problem that there is no loss of generality in assuming that both f,g are Borel measurable. Next show this makes sense a.e. $\boldsymbol{x}$ and that in fact for a.e. $\boldsymbol{x}$ $\int|f(\boldsymbol{x}-\boldsymbol{y})||g(\boldsymbol{y})|dm_p(y)<\infty$. Next show

$$\int|f*g(\boldsymbol{x})|dm_p(\boldsymbol{x})\le\int|f|dm_p\int|g|dm_p.$$

Hint: You can use Fubini's theorem to write

$$\int\int|f(\boldsymbol{x}-\boldsymbol{y})||g(\boldsymbol{y})|dm_p(y)dm_p(x)=$$

$$\int\int|f(\boldsymbol{x}-\boldsymbol{y})||g(\boldsymbol{y})|dm_p(x)dm_p(y)=\int|f(\boldsymbol{z})|dm_p\int|g(\boldsymbol{y})|dm_p.$$

11. Suppose $X:(\Omega,\mathscr{F},P)$ where P is a probability measure and suppose $X:\Omega\to\mathbb{R}$ is measurable. That is, $X^{-1}(\text{open set})\in\mathscr{F}$. Then consider the distribution measure λ_X defined on the Borel sets of $\mathbb{R}^p$ and given as follows. $\lambda_X(E)=P(X\in E)$. Explain why this is a probability measure on $\mathscr{B}(\mathbb{R})$ and why $X^{-1}(B)\in\mathscr{F}$ whenever B is a Borel set. Next show that if $X\in L^1(\Omega)$, $\int_\Omega XdP=\int_{\mathbb{R}}xd\lambda_X$. Suppose h is a complex valued Borel measurable function defined on $\mathbb{R}$ which is also bounded. Show that

$$\int h(x)d\lambda_X(x)=\int h(X(\omega))dP$$

Hint: Recall that from the definition of the integral,

$$\int_{\mathbb{R}}|x|d\lambda_X=\int_0^\infty\lambda_X(|x|>\alpha)d\alpha=\int_0^\infty P(|X|>\alpha)d\alpha=\int_\Omega|X|dP<\infty$$

12. Let $V = \boldsymbol{h}(U)$ where U, V are open in $\mathbb{R}^p$ and $\boldsymbol{h}$ and $\boldsymbol{h}^{-1}$ are one to one and differentiable. Show without using the Brouwer fixed point theorem that $\int_V f(\boldsymbol{y})\, dm_p = \int_U f(\boldsymbol{h}(\boldsymbol{x}))\, |\det D\boldsymbol{h}(\boldsymbol{x})|\, dm_p$. **Hint:** Estimate $D_{m_p}(\lambda)(\boldsymbol{x})$ showing that it is no more than $|\det D\boldsymbol{h}(\boldsymbol{x})|$ and then use what you just showed on $\boldsymbol{h}^{-1} : V \to U$ and chain rule. Note that if $V = \boldsymbol{h}(U)$, then by invariance of domain, V will be open.

13. Let $\boldsymbol{h} : U \to \boldsymbol{h}(U)$ be one to one and C^1. Show $\{\boldsymbol{x} : |\det D\boldsymbol{h}(\boldsymbol{x})| = 0\}$ is closed. Then apply the inverse function theorem and Sard's lemma with what was just shown above. These two problems don't require the Brouwer fixed point theorem. Now obtain Theorem 11.16.3.

14. The next few problems involve a review of the invariance of domain theorem presented earlier. Suppose U is a nonempty open set in $\mathbb{R}^p$, $\boldsymbol{f} : U \to \mathbb{R}^p$ is continuous, and suppose that for each $\boldsymbol{x} \in U$, there is a ball $B_{\boldsymbol{x}}$ containing $\boldsymbol{x}$ such that $\boldsymbol{f}$ is one to one on $B_{\boldsymbol{x}}$. That is, $\boldsymbol{f}$ is locally one to one. Show that $\boldsymbol{f}(U)$ is open.

15. $\uparrow$ In the situation of the above problem, suppose $\boldsymbol{f} : \mathbb{R}^p \to \mathbb{R}^p$ is locally one to one. Also suppose that $\lim_{|\boldsymbol{x}|\to\infty} |\boldsymbol{f}(\boldsymbol{x})| = \infty$. Show that it follows that $\boldsymbol{f}(\mathbb{R}^p) = \mathbb{R}^p$. That is, $\boldsymbol{f}$ is onto. Show that this would not be true if $\boldsymbol{f}$ is only defined on a proper open set. Also show that this would not be true if the condition $\lim_{|\boldsymbol{x}|\to\infty} |\boldsymbol{f}(\boldsymbol{x})| = \infty$ does not hold. **Hint:** You might show that $\boldsymbol{f}(\mathbb{R}^p)$ is both open and closed and then use connectedness. To get an example in the second case, you might think of e^{x+iy}. It does not include $0 + i0$. Why not?

16. $\uparrow$ Show that if $\boldsymbol{f} : \mathbb{R}^p \to \mathbb{R}^p$ is C^1 and if $D\boldsymbol{f}(\boldsymbol{x})$ exists and is invertible for all $\boldsymbol{x} \in \mathbb{R}^p$, then $\boldsymbol{f}$ is locally one to one. Thus, from the above problem, if $\lim_{|\boldsymbol{x}|\to\infty} |\boldsymbol{f}(\boldsymbol{x})| = \infty$, then $\boldsymbol{f}$ is also onto.

17. If a continuous function is one to one on a compact set, explain why its inverse is continuous.

18. You know from linear algebra that there is no onto linear mapping $A : \mathbb{R}^m \to \mathbb{R}^p$ for $p > m$. Show that there is no locally one to one continuous mapping which will map $\mathbb{R}^m$ onto $\mathbb{R}^p$.

19. Suppose U is a nonempty set in $\mathbb{R}^p$. Let ∂U consist of the points $\boldsymbol{p} \in \mathbb{R}^p$ such that $B(\boldsymbol{p}, r)$ contains points of U as well as points of $\mathbb{R}^p \setminus U$. Show that U is contained in the union of the interior of U, denoted as $\text{int}(U)$ with ∂U. Now suppose that $\boldsymbol{f} : U \to \mathbb{R}^p$ and is one to one and continuous. Explain why $\text{int}(\boldsymbol{f}(U))$ equals $\boldsymbol{f}(\text{int}(U))$.

20. Prove the Radon Nikodym theorem, Theorem 11.6.2 in case $\lambda \ll \mu$ another way by using the earlier general Radon Nikodym theorem, Theorem 10.13.7 or its corollary and then identifying the function ensured by that theorem with the symmetric derivative, using the fundamental theorem of calculus, Theorem 11.3.2.

21. Suppose $\boldsymbol{x}$ is a Lebesgue point of f with respect to Lebesgue measure so that

$$\lim_{r\to 0} \frac{1}{m_p(\boldsymbol{x}, r)} \int_{B(\boldsymbol{x},r)} |f(\boldsymbol{x}) - f(\boldsymbol{y})|\, dm_p(y) = 0.$$

Suppose $\boldsymbol{x} \in E_r \subseteq B(\boldsymbol{x}, \sigma r)$, E_r a measurable set and there exists δ such that

$$\delta m_p(B(\boldsymbol{x}, r)) < m_p(E_r)$$

for all r. Verify that $\lim_{r\to 0}\frac{1}{m_p(E_r)}\int_{E_r}|f(\boldsymbol{x})-f(\boldsymbol{y})|\,dm_p(y)=0$.

22. Generalize Corollary 11.7.6 to the case where m is replaced by m_p and μ is some Radon measure on $\mathscr{F}_\mu$ a σ algebra of sets of $\mathbb{R}^p$ or an open subset of $\mathbb{R}^p$.

23. If F is increasing and continuous and I consists of all t such that $F'_+(t)=\infty$ or $F'_-(t)=\infty$, show that if $\mu(I)=0$ for μ the Lebesgue Stieltjes measure associated with F, then F is absolutely continuous.

24. Let

$$I\equiv\left\{t:\min\left(\liminf_{h\to 0}\frac{F(t+h)-F(t)}{h},\liminf_{h\to 0}\frac{F(t)-F(t-h)}{h}\right)=\infty\right\}$$

where F is increasing and continuous. Let μ be the Lebesgue Stieltjes measure coming from F. Show that if $\mu(I)=0$, then $\mu\ll m$. This I is the set where $F'(t)=\infty$. Conversely, if $\mu\ll m$, then F is continuous and $\mu(I)=\infty$. **Hint:** Let $G_n\equiv\left\{t:\min\left(\liminf_{h\to 0}\frac{F(t+h)-F(t)}{h},\liminf_{h\to 0}\frac{F(t)-F(t-h)}{h}\right)\le n\right\}$ and follow the idea of Corollary 11.7.6 using a covering theorem, Theorem 9.12.2. The assumption that F is continuous is needed to say that, for example, $\frac{F(t+h)-F(t)}{h}=\frac{\mu([t,t+h])}{h}$.

Chapter 12

The L^p Spaces

12.1 Basic Inequalities and Properties

One of the main applications of the Lebesgue integral is to the study of various sorts of functions space. These are vector spaces whose elements are functions of various types. One of the most important examples of a function space is the space of measurable functions whose absolute values are p^{th} power integrable where $p \geq 1$. These spaces, referred to as L^p spaces, are very useful in applications. In the chapter $(\Omega, \mathscr{S}, \mu)$ will be a measure space.

Definition 12.1.1 *Let $1 \leq p < \infty$. Define*

$$L^p(\Omega) \equiv \{f : f \text{ is measurable and } \int_\Omega |f(\omega)|^p d\mu < \infty\}$$

In terms of the distribution function,

$$L^p(\Omega) = \{f : f \text{ is measurable and } \int_0^\infty pt^{p-1}\mu([|f| > t])\, dt < \infty\}$$

because $\int_0^\infty \mu([|f|^p > t])\, dt = \int_0^\infty \mu\left(|f| > t^{1/p}\right) dt = \int_0^\infty ps^{p-1}\mu([|f| > s])\, ds$.

For each $p > 1$ define q by $\frac{1}{p} + \frac{1}{q} = 1$. Often one uses p' instead of q in this context.

$L^p(\Omega)$ is a vector space and has a norm. This is similar to the situation for $\mathbb{R}^n$ but the proof requires the following fundamental inequality. When $p = 1$, we use the symbol ∞ to represent q. The space $L^\infty(\Omega)$ will be discussed later.

Theorem 12.1.2 *(Holder's inequality) If f and g are measurable functions, then if $p > 1$,*

$$\int |f|\,|g|\, d\mu \leq \left(\int |f|^p d\mu\right)^{\frac{1}{p}} \left(\int |g|^q d\mu\right)^{\frac{1}{q}}. \tag{12.1}$$

Proof: First recall Lemma 4.3.10, stated here for convenience.

Lemma 12.1.3 *If $p > 1$, and $0 \leq a, b$ then $ab \leq \frac{a^p}{p} + \frac{b^q}{q}$. Equality occurs when $a^p = b^q$.*

Proof of Holder's inequality: If either $\int |f|^p d\mu$ or $\int |g|^p d\mu$ equals ∞, the inequality 12.1 is obviously valid because $\infty \geq$ anything. If either $\int |f|^p d\mu$ or $\int |g|^p d\mu$ equals 0, then $f = 0$ a.e. or $g = 0$ a.e. and so in this case the left side of the inequality equals 0 and so the inequality is therefore true. Therefore assume both $\int |f|^p d\mu$ and $\int |g|^p d\mu$ are less than ∞ and not equal to 0. Let $\left(\int |f|^p d\mu\right)^{1/p} = I(f)$ and let $\left(\int |g|^p d\mu\right)^{1/q} = I(g)$. Then using the lemma,

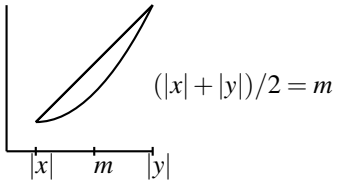
$$\int \frac{|f|}{I(f)} \frac{|g|}{I(g)}\, d\mu \leq \frac{1}{p}\int \frac{|f|^p}{I(f)^p}\, d\mu + \frac{1}{q}\int \frac{|g|^q}{I(g)^q}\, d\mu = 1.$$

Hence, $\int |f|\,|g|\, d\mu \leq I(f)I(g) = \left(\int |f|^p d\mu\right)^{1/p}\left(\int |g|^q d\mu\right)^{1/q}$. This proves Holder's inequality. ■

The following lemma will be needed.

Lemma 12.1.4 *Suppose $x, y \in \mathbb{C}$. Then $|x+y|^p \leq 2^{p-1}\left(|x|^p + |y|^p\right)$.*

Proof: The function $f(t) = t^p$ is concave up for $t \geq 0$ because $p > 1$. Therefore, the secant line joining two points on the graph of this function must lie above the graph of the function. This is illustrated in the following picture.



Since $\left(\frac{|x|+|y|}{2}\right)^p \leq \frac{|x|^p+|y|^p}{2}$, $|x+y|^p \leq (|x|+|y|)^p \leq 2^{p-1}\left(|x|^p + |y|^p\right)$ ■

Note that if $y = \phi(x)$ is any function for which the graph of ϕ is concave up, you could get a similar inequality by the same argument.

Corollary 12.1.5 *(Minkowski inequality) Let $1 \leq p < \infty$. Then*

$$\left(\int |f+g|^p \, d\mu\right)^{1/p} \leq \left(\int |f|^p \, d\mu\right)^{1/p} + \left(\int |g|^p \, d\mu\right)^{1/p}. \tag{12.2}$$

Proof: If $p = 1$, this is obvious. Let $p > 1$. Without loss of generality, assume $\left(\int |f|^p \, d\mu\right)^{1/p} + \left(\int |g|^p \, d\mu\right)^{1/p} < \infty$ and $\left(\int |f+g|^p \, d\mu\right)^{1/p} \neq 0$ or there is nothing to prove. Therefore, using the above lemma,

$$\int |f+g|^p d\mu \leq 2^{p-1}\left(\int |f|^p + |g|^p d\mu\right) < \infty.$$

Now $|f(\omega)+g(\omega)|^p \leq |f(\omega)+g(\omega)|^{p-1}\left(|f(\omega)|+|g(\omega)|\right)$. Also, it follows from the definition of p and q that $p-1 = \frac{p}{q}$. Therefore, using this and Holder's inequality, $\int |f+g|^p d\mu \leq$

$$\begin{aligned} & \int |f+g|^{p-1}|f| d\mu + \int |f+g|^{p-1}|g| d\mu = \int |f+g|^{\frac{p}{q}}|f| d\mu + \int |f+g|^{\frac{p}{q}}|g| d\mu \\ \leq \quad & \left(\int |f+g|^p d\mu\right)^{\frac{1}{q}}\left(\int |f|^p d\mu\right)^{\frac{1}{p}} + \left(\int |f+g|^p d\mu\right)^{\frac{1}{q}}\left(\int |g|^p d\mu\right)^{\frac{1}{p}}. \end{aligned}$$

Dividing both sides by $\left(\int |f+g|^p d\mu\right)^{\frac{1}{q}}$ yields 12.2. ■

The above theorem implies the following corollary.

Corollary 12.1.6 *Let $f_i \in L^p(\Omega)$ for $i = 1, 2, \cdots, n$. Then*

$$\left(\int \left|\sum_{i=1}^n f_i\right|^p d\mu\right)^{1/p} \leq \sum_{i=1}^n \left(\int |f_i|^p\right)^{1/p}.$$

This shows that if $f, g \in L^p$, then $f+g \in L^p$. Also, it is clear that if a is a constant and $f \in L^p$, then $af \in L^p$ because $\int |af|^p \, d\mu = |a|^p \int |f|^p \, d\mu < \infty$. Thus L^p is a vector space and

a.) $\left(\int |f|^p \, d\mu\right)^{1/p} \geq 0, \left(\int |f|^p \, d\mu\right)^{1/p} = 0$ if and only if $f = 0$ a.e.

b.) $\left(\int |af|^p \, d\mu\right)^{1/p} = |a| \left(\int |f|^p \, d\mu\right)^{1/p}$ if a is a scalar.

c.) $\left(\int |f+g|^p \, d\mu\right)^{1/p} \leq \left(\int |f|^p \, d\mu\right)^{1/p} + \left(\int |g|^p \, d\mu\right)^{1/p}$.

$f \to \left(\int |f|^p \, d\mu\right)^{1/p}$ would define a norm if $\left(\int |f|^p \, d\mu\right)^{1/p} = 0$ implied $f = 0$. Unfortunately, this is not so because if $f = 0$ a.e. but is nonzero on a set of measure zero, $\left(\int |f|^p \, d\mu\right)^{1/p} = 0$ and this is not allowed. However, all the other properties of a norm are available and so a little thing like a set of measure zero will not prevent the consideration of L^p as a normed vector space if two functions in L^p which differ only on a set of measure zero are considered the same. That is, an element of L^p is really an equivalence class of functions where two functions are equivalent if they are equal a.e. With this convention, here is a definition.

Definition 12.1.7 *Let $f \in L^p(\Omega)$. Define $\|f\|_p \equiv \|f\|_{L^p} \equiv \left(\int |f|^p \, d\mu\right)^{1/p}$.*

Then with this definition and using the convention that elements in L^p are considered to be the same if they differ only on a set of measure zero, $\| \|_p$ is a norm on $L^p(\Omega)$ because if $||f||_p = 0$ then $f = 0$ a.e. and so f is considered to be the zero function because it differs from 0 only on a set of measure zero.

The following is an important definition.

Definition 12.1.8 *A complete normed linear space is called a Banach[1] space.*

L^p is a Banach space. This is the next big theorem which says that these L^p spaces are always complete.

Theorem 12.1.9 *The following holds for $L^p(\Omega, \mathscr{F}, \mu), p \geq 1$. If $\{f_n\}$ is a Cauchy sequence in $L^p(\Omega)$, then there exists $f \in L^p(\Omega)$ and a subsequence which converges a.e. to $f \in L^p(\Omega)$, and $\|f_n - f\|_p \to 0$.*

Proof: Let $\{f_n\}$ be a Cauchy sequence in $L^p(\Omega)$. This means that for every $\varepsilon > 0$ there exists N such that if $n, m \geq N$, then $\|f_n - f_m\|_p < \varepsilon$. Now select a subsequence as follows. Let n_1 be such that $\|f_n - f_m\|_p < 2^{-1}$ whenever $n, m \geq n_1$. Let n_2 be such that $n_2 > n_1$ and $\|f_n - f_m\|_p < 2^{-2}$ whenever $n, m \geq n_2$. If $n_1, \cdots, n_k$ have been chosen, let $n_{k+1} > n_k$ and whenever $n, m \geq n_{k+1}, \|f_n - f_m\|_p < 2^{-(k+1)}$. The subsequence just mentioned is $\{f_{n_k}\}$. Thus

$$\mu\left(\left\{\omega : \left|f_{n_k}(\omega) - f_{n_{k+1}}(\omega)\right|^p > \left(\frac{2}{3}\right)^k\right\}\right) \equiv \mu(E_k)$$

$$\leq \left(\frac{3}{2}\right)^k \int_{E_k} \left|f_{n_k}(\omega) - f_{n_{k+1}}(\omega)\right|^p d\mu$$

[1] These spaces are named after Stefan Banach, 1892-1945. Banach spaces are the basic item of study in the subject of functional analysis and will be considered later in this book.

There is a recent biography of Banach, R. Katuża, *The Life of Stefan Banach,* (A. Kostant and W. Woyczyński, translators and editors) Birkhauser, Boston (1996). More information on Banach can also be found in a recent short article written by Douglas Henderson who is in the department of chemistry and biochemistry at BYU.

Banach was born in Austria, worked in Poland and died in the Ukraine but never moved. This is because borders kept changing. There is a rumor that he died in a German concentration camp which is apparently not true. It seems he died after the war of lung cancer.

He was an interesting character. He hated taking examinations so much that he did not receive his undergraduate university degree. Nevertheless, he did become a professor of mathematics due to his important research. He and some friends would meet in a cafe called the Scottish cafe where they wrote on the marble table tops until Banach's wife supplied them with a notebook which became the "Scotish notebook" and was eventually published.

$$\leq \left(\frac{3}{2}\right)^k \|f_{n_k} - f_{n_{k+1}}\|_{L^p}^p < \left(\frac{3}{2}\right)^k 2^{-kp} \leq \frac{3^k}{4^k}$$

Hence $\sum_k \mu(E_k) < \infty$ and so, by the Borel Cantelli lemma, Lemma 9.2.5 on Page 247, there is a set of measure zero N such that if $\omega \notin N$, then $\left|f_{n_k}(\omega) - f_{n_{k+1}}(\omega)\right| \leq \left(\frac{2}{3}\right)^{k/p}$. Since

$$\sum_k \left(\frac{2}{3}\right)^{k/p} < \infty, \left\{\mathcal{X}_{N^C}(\omega) f_{n_k}(\omega)\right\}_{k=1}^{\infty}$$

is a Cauchy sequence for all ω. Let it converge to $f(\omega)$, a measurable function since it is a limit of measurable functions. By Fatou's lemma, and the Minkowski inequality, Corollary 12.1.5, $\|f - f_{n_k}\|_p = \left(\int |f - f_{n_k}|^p \, d\mu\right)^{1/p} \leq$

$$\liminf_{m\to\infty} \left(\int |f_{n_m} - f_{n_k}|^p \, d\mu\right)^{1/p} = \liminf_{m\to\infty} \|f_{n_m} - f_{n_k}\|_p \leq$$

$$\liminf_{m\to\infty} \sum_{j=k}^{m-1} \left\|f_{n_{j+1}} - f_{n_j}\right\|_p \leq \sum_{i=k}^{\infty} \|f_{n_{i+1}} - f_{n_i}\|_p \leq 2^{-(k-1)}. \tag{12.3}$$

Therefore, $f \in L^p(\Omega)$ because $\|f\|_p \leq \|f - f_{n_k}\|_p + \|f_{n_k}\|_p < \infty$, and $\lim_{k\to\infty} \|f_{n_k} - f\|_p = 0$. This proves b.).

This has shown f_{n_k} converges to f in $L^p(\Omega)$. It follows the original Cauchy sequence also converges to f in $L^p(\Omega)$. This is a general fact that if a subsequence of a Cauchy sequence converges, then so does the original Cauchy sequence. This is Theorem 3.2.2. ■

In working with the L^p spaces, the following inequality also known as Minkowski's inequality is very useful. See [25]. It is similar to the Minkowski inequality for sums. To see this, replace the integral, $\int_X$ with a finite summation sign and you will see the usual Minkowski inequality or rather the version of it given in Corollary 12.1.6.

Lemma 12.1.10 *Let $(X, \mathcal{S}, \mu)$ and $(Y, \mathcal{F}, \lambda)$ be finite measure spaces and let f be $\mu \times \lambda$ measurable. Then the following inequality is valid for $p \geq 1$.*

$$\int_X \left(\int_Y |f(x,y)|^p \, d\lambda\right)^{\frac{1}{p}} d\mu \geq \left(\int_Y \left(\int_X |f(x,y)| \, d\mu\right)^p d\lambda\right)^{\frac{1}{p}}. \tag{12.4}$$

Proof: This is an application of the Fubini theorem and Holder inequality. Recall that $p - 1 = p/p'$. Let $J(y) \equiv \int_X |f(x,y)| \, d\mu$. Then

$$\int_Y \left(\int_X |f(x,y)| \, d\mu\right)^p d\lambda = \int_Y J(y)^{p/p'} \int_X |f(x,y)| \, d\mu d\lambda$$

$$= \int_Y \int_X |f(x,y)| J(y)^{p/p'} \, d\mu d\lambda = \int_X \int_Y |f(x,y)| J(y)^{p/p'} \, d\lambda d\mu$$

$$\leq \left(\int_Y J(y)^p \, d\lambda\right)^{1/p'} \int_X \left(\int_Y |f(x,y)|^p \, d\lambda\right)^{1/p} d\mu$$

$$= \left(\int_Y \left(\int_X |f(x,y)| \, d\mu\right)^p d\lambda\right)^{1/p'} \int_X \left(\int_Y |f(x,y)|^p \, d\lambda\right)^{1/p} d\mu$$

Thus $\int_Y \left(\int_X |f(x,y)|\,d\mu\right)^p d\lambda \leq$

$$\left(\int_Y \left(\int_X |f(x,y)|\,d\mu\right)^p d\lambda\right)^{1/p'} \int_X \left(\int_Y |f(x,y)|^p\,d\lambda\right)^{1/p} d\mu \tag{12.5}$$

If f is bounded, divide both sides by the first factor on the right and obtain 12.4. Otherwise replace f with $\min(f,n)$, divide and then apply the monotone convergence theorem as $n \to \infty$ to get 12.4. Note that 12.4 holds even if the first factor on the right in 12.5 equals zero. ■

Now consider the case where the measure spaces are σ finite.

Theorem 12.1.11 *Let $(X,\mathcal{S},\mu)$ and $(Y,\mathcal{F},\lambda)$ be σ-finite measure spaces and let f be product measurable. Then the following inequality is valid for $p \geq 1$.*

$$\int_X \left(\int_Y |f(x,y)|^p\,d\lambda\right)^{\frac{1}{p}} d\mu \geq \left(\int_Y \left(\int_X |f(x,y)|\,d\mu\right)^p d\lambda\right)^{\frac{1}{p}}. \tag{12.6}$$

Proof: Since the two measure spaces are σ finite, there exist measurable sets, X_m and Y_k such that $X_m \subseteq X_{m+1}$ for all m, $Y_k \subseteq Y_{k+1}$ for all k, and $\mu(X_m), \lambda(Y_k) < \infty$. From the above,

$$\int_{X_m} \left(\int_{Y_k} |f(x,y)|^p\,d\lambda\right)^{\frac{1}{p}} d\mu \geq \left(\int_{Y_k} \left(\int_{X_m} |f(x,y)|\,d\mu\right)^p d\lambda\right)^{\frac{1}{p}}. \tag{12.7}$$

Now use the monotone convergence theorem to pass to a limit first as $k \to \infty$ and then as $m \to \infty$. ■

Note that the proof of this theorem depends on two manipulations, the interchange of the order of integration and Holder's inequality. Also observe that there is nothing to check in the case of double sums. Thus if $a_{ij} \geq 0$, it is always the case that $\left(\sum_j \left(\sum_i a_{ij}\right)^p\right)^{1/p} \leq \sum_i \left(\sum_j a_{ij}^p\right)^{1/p}$ because the integrals in this case are just sums and $(i,j) \to a_{ij}$ is measurable.

The L^p spaces have many important properties. Before considering these, here is a definition of L^∞

Definition 12.1.12 *$f \in L^\infty(\Omega,\mu)$ if there exists a set of measure zero E, and a constant $C < \infty$ such that $|f(x)| \leq C$ for all $x \notin E$. Then $\|f\|_\infty$ is defined as $\|f\|_\infty \equiv \inf\{C : |f(x)| \leq C \text{ a.e.}\}$, the* inf *of all such C.*

Proposition 12.1.13 *$\|\ \|_\infty$ is a norm on $L^\infty(\Omega,\mu)$ provided f and g are identified if $f(x) = g(x)$ a.e. Also, $L^\infty(\Omega,\mu)$ is complete. In addition to this, $\|f\|_\infty$ has the property that $|f(x)| \leq \|f\|_\infty$ for a.e. x so the norm is the smallest constant with this property.*

Proof: It is obvious that $\|f\|_\infty \geq 0$. Let $C_n \downarrow \|f\|_\infty$ where $|f(x)| \leq C_n$ off a set of measure zero E_n. Then let $E \equiv \cup_n E_n$. This is also a set of measure zero and if $x \notin E$, then $|f(x)| \leq C_n$ for all C_n and so $|f(x)| \leq \|f\|_\infty$ for all $x \notin E$. Thus $\|f\|_\infty$ is the smallest number C with $|f(x)| \leq C$ a.e. In case $\|f\|_\infty = 0$, $f(x) = 0$ off of E and so we regard f as 0 because it equals 0 a.e.

If $c = 0$ there is nothing to show in the claim that $\|cf\|_\infty = |c|\,\|f\|_\infty$. Assume then that $c \neq 0$. $|c|\,\|f\|_\infty \geq |cf(x)|$ a.e. Thus $|c|\,\|f\|_\infty \geq \|cf\|_\infty$ whenever $c \neq 0$. Thus $\|f\|_\infty = \left\|\frac{1}{c}cf\right\|_\infty \leq \frac{1}{|c|}\|cf\|_\infty$ and so $|c|\,\|f\|_\infty \leq \|cf\|_\infty$. It remains to verify the triangle inequality.

It was just shown that $\|f\|_\infty$ is the smallest constant such that $|f(x)| \leq \|f\|_\infty$ a.e. Hence if $f, g \in L^\infty(\Omega)$, $|f(x)+g(x)| \leq |f(x)|+|g(x)| \leq \|f\|_\infty + \|g\|_\infty$ a.e. and so by definition $\|f+g\|_\infty \leq \|f\|_\infty + \|g\|_\infty$.

Next suppose you have a Cauchy sequence in $L^\infty(\Omega)$ $\{f_n\}$. Let $|f_n(x)-f_m(x)| < \|f_n-f_m\|_\infty$ for $x \notin E_{nm}$, $\mu(E_{nm})=0$ and let $|f_n(x)| \leq \|f_n\|_\infty$ for $x \notin E_n$, $\mu(E_n)=0$. Then let $E \equiv \cup_n E_n \cup \cup_{m,n} E_{nm}$. It follows that for $x \notin E$, $\lim_{n\to\infty} f_n(x)$ exists. Let $f(x)$ be this limit for $x \notin E$ and let $f(x)=0$ on E. Also $|\|f_n\|_\infty - \|f_m\|_\infty| \leq \|f_n-f_m\|_\infty$ since $\|\cdot\|_\infty$ is a norm. Therefore, for $x \notin E$, $|f(x)| = \lim_{n\to\infty} |f_n(x)| \leq \lim_{n\to\infty} \|f_n\|_\infty \equiv C$ so $f \in L^\infty(\Omega,\mu)$. Also, for $x \notin E$, $|f_m(x)-f_n(x)| \leq \|f_m-f_n\|_\infty < \varepsilon$ if $m > n$ and n is large enough. Therefore, for such n, letting $m \to \infty$, $|f(x)-f_n(x)| \leq \varepsilon$ for $x \notin E$. It follows that $\|f-f_n\|_\infty \leq \varepsilon$ if n large enough and so by definition, $\lim_{n\to\infty} \|f-f_n\|_\infty = 0$. ■

12.2 Density Considerations

Theorem 12.2.1 *Let $p \geq 1$ and let $(\Omega, \mathscr{S}, \mu)$ be a measure space. Then the simple functions are dense in $L^p(\Omega)$. In fact, if $f \in L^p(\Omega)$, then there is a sequence of simple functions $\{s_n\}$ such that $|s_n| \leq |f|$ and $\|f-s_n\|_p \to 0$.*

Proof: Recall that a function f, having values in $\mathbb{R}$ can be written in the form $f = f^+ - f^-$ where

$$f^+ = \max(0,f), \ f^- = \max(0,-f).$$

Therefore, an arbitrary complex valued function, f is of the form

$$f = \mathrm{Re}\, f^+ - \mathrm{Re}\, f^- + i\left(\mathrm{Im}\, f^+ - \mathrm{Im}\, f^-\right).$$

If each of these nonnegative functions is approximated by a simple function, it follows f is also approximated by a simple function. Approximating each of the positive and negative parts with simple functions having absolute value less than what is approximated, it would follow that $|s_n| \leq 4|f|$ and all that is left is to verify that $\|s_n - f\|_p \to 0$ which occurs if it happens for each of these positive and negative parts of real and imaginary parts.

Now $|f(x)-s_n(x)| \leq 5|f|$ and so $|f(x)-s_n(x)|^p \leq 5^p |f|^p$ which is in L^1. Then by the dominated convergence theorem, $0 = \lim_{n\to\infty} \int |f(x)-s_n(x)|^p d\mu$ showing that the simple functions are dense in L^p. ■

Note how this observation always holds and requires no assumptions on the measures.

Recall that for Ω a topological space, $C_c(\Omega)$ is the space of continuous functions with compact support in Ω. Also recall the following definition.

Definition 12.2.2 *Let $(\Omega, \mathscr{S}, \mu)$ be a measure space and suppose (Ω, τ) is also a topological space (metric space if you like.). Then $(\Omega, \mathscr{S}, \mu)$ is called a regular measure space if the σ algebra of Borel sets is contained in $\mathscr{S}$ and for all $E \in \mathscr{S}$,*

$$\mu(E) = \inf\{\mu(V) : V \supseteq E \text{ and } V \text{ open}\}$$

and if $\mu(E) < \infty$,

$$\mu(E) = \sup\{\mu(K) : K \subseteq E \text{ and } K \text{ is compact }\}$$

and $\mu(K) < \infty$ for any compact set, K.

For example Lebesgue measure is an example of such a measure. More generally these measures are often referred to as Radon measures when they are complete. Recall the following important result which is Lemma 3.12.4.

Lemma 12.2.3 *Let Ω be a metric space in which the closed balls are compact and let K be a compact subset of V, an open set. Then there exists a continuous function $f : \Omega \to [0,1]$ such that $f(x) = 1$ for all $x \in K$ and $\mathrm{spt}(f)$ is a compact subset of V. That is, $K \prec f \prec V$.*

It is not necessary to be in a metric space to do this. You can accomplish the same thing using Urysohn's lemma in a normal topological space or, as is often done, a locally compact Hausdorff space. This can be discussed later.

Theorem 12.2.4 *Let $(\Omega, \mathscr{S}, \mu)$ be a regular measure space as in Definition 12.2.2 where the conclusion of Lemma 3.12.4 holds. Then $C_c(\Omega)$ is dense in $L^p(\Omega)$.*

Proof: First consider a measurable set, E where $\mu(E) < \infty$. Let $K \subseteq E \subseteq V$ where $\mu(V \setminus K) < \varepsilon$. Now let $K \prec h \prec V$. Then

$$\int |h - \mathscr{X}_E|^p \, d\mu \leq \int \mathscr{X}_{V \setminus K}^p d\mu = \mu(V \setminus K) < \varepsilon.$$

It follows that for each s a simple function in $L^p(\Omega)$, there exists $h \in C_c(\Omega)$ such that $\|s - h\|_p < \varepsilon$. This is because if $s(x) = \sum_{i=1}^m c_i \mathscr{X}_{E_i}(x)$ is a simple function in L^p where the c_i are the distinct nonzero values of s each $\mu(E_i) < \infty$ since otherwise $s \notin L^p$ due to the inequality $\int |s|^p \, d\mu \geq |c_i|^p \mu(E_i)$. By Theorem 12.2.1, simple functions are dense in $L^p(\Omega)$. Therefore, $C_c(\Omega)$ is dense in $L^p(\Omega)$. ■

12.3 Separability

The most important case is of course Lebesgue measure on $\mathbb{R}^n$ or more generally, some Radon measure.

Theorem 12.3.1 *For $p \geq 1$ and μ a Radon measure, $L^p(\mathbb{R}^n, \mu)$ is separable. Recall this means there exists a countable set $\mathscr{D}$ such that if $f \in L^p(\mathbb{R}^n, \mu)$ and $\varepsilon > 0$, there exists $g \in \mathscr{D}$ such that $\|f - g\|_p < \varepsilon$.*

Proof: Let Q be all functions of the form $c\mathscr{X}_{[\boldsymbol{a},\boldsymbol{b})}$ where

$$[\boldsymbol{a},\boldsymbol{b}) \equiv [a_1, b_1) \times [a_2, b_2) \times \cdots \times [a_n, b_n),$$

and both a_i, b_i are rational, while c has rational real and imaginary parts. Let $\mathscr{D}$ be the set of all finite sums of functions in Q. Thus, $\mathscr{D}$ is countable. In fact $\mathscr{D}$ is dense in $L^p(\mathbb{R}^n, \mu)$. To prove this, it is necessary to show that for every $f \in L^p(\mathbb{R}^n, \mu)$, there exists an element of $\mathscr{D}$, s such that $\|s - f\|_p < \varepsilon$. If it can be shown that for every $g \in C_c(\mathbb{R}^n)$ there exists $h \in \mathscr{D}$ such that $\|g - h\|_p < \varepsilon$, then this will suffice because if $f \in L^p(\mathbb{R}^n)$ is arbitrary, Theorem 12.2.4 implies there exists $g \in C_c(\mathbb{R}^n)$ such that $\|f - g\|_p \leq \frac{\varepsilon}{2}$ and then there would exist $h \in C_c(\mathbb{R}^n)$ such that $\|h - g\|_p < \frac{\varepsilon}{2}$. By the triangle inequality,

$$\|f - h\|_p \leq \|h - g\|_p + \|g - f\|_p < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Therefore, assume at the outset that $f \in C_c(\mathbb{R}^n)$.

Let $\mathscr{P}_m$ consist of all sets of the form $[\boldsymbol{a},\boldsymbol{b}) \equiv \prod_{i=1}^n [a_i,b_i)$ where $a_i = j2^{-m}$and $b_i = (j+1)2^{-m}$ for j an integer. Thus $\mathscr{P}_m$ consists of a tiling of $\mathbb{R}^n$ into half open rectangles having diameters $2^{-m}n^{\frac{1}{2}}$. There are countably many of these rectangles; so let$\mathscr{P}_m = \{[\boldsymbol{a}_i,\boldsymbol{b}_i)\}$ for $i \geq 1$, and $\mathbb{R}^n = \cup_{i=1}^\infty [\boldsymbol{a}_i,\boldsymbol{b}_i)$. Let c_i^m be complex numbers with rational real and imaginary parts satisfying

$$|f(\boldsymbol{a}_i) - c_i^m| < 2^{-m},\ |c_i^m| \leq |f(\boldsymbol{a}_i)|. \tag{12.8}$$

Let $s_m(\boldsymbol{x}) = \sum_{i=1}^\infty c_i^m \mathscr{X}_{[\boldsymbol{a}_i,\boldsymbol{b}_i)}(\boldsymbol{x})$.

Since $f(\boldsymbol{a}_i) = 0$ except for finitely many values of i, the above is a finite sum. Then 12.8 implies $s_m \in \mathscr{D}$. If s_m converges uniformly to f then it will follow that s_m is close to f in L^p.

Since $f \in C_c(\mathbb{R}^n)$ it follows that f is uniformly continuous and so given $\varepsilon > 0$ there exists $\delta > 0$ such that if $|\boldsymbol{x}-\boldsymbol{y}| < \delta$, $|f(\boldsymbol{x}) - f(\boldsymbol{y})| < \varepsilon/2$. Now let m be large enough that every box in $\mathscr{P}_m$ has diameter less than δ and also that $2^{-m} < \varepsilon/2$. Then if $[\boldsymbol{a}_i,\boldsymbol{b}_i)$ is one of these boxes of $\mathscr{P}_m$, and $\boldsymbol{x} \in [\boldsymbol{a}_i,\boldsymbol{b}_i)$, $|f(\boldsymbol{x}) - f(\boldsymbol{a}_i)| < \varepsilon/2$ and $|f(\boldsymbol{a}_i) - c_i^m| < 2^{-m} < \varepsilon/2$. Therefore, using the triangle inequality, it follows that for $\boldsymbol{x} \in [\boldsymbol{a}_i,\boldsymbol{b}_i)$,

$$\begin{aligned} |f(\boldsymbol{x}) - s_m(\boldsymbol{x})| &= |f(\boldsymbol{x}) - c_i^m| = |f(\boldsymbol{x}) - f(\boldsymbol{a}_i)| + |f(\boldsymbol{a}_i) - c_i^m| \\ &< 2^{-m} + 2^{-m} < \varepsilon \end{aligned}$$

and since $\boldsymbol{x}$ is arbitrary, this establishes uniform convergence. From the construction, s_m and f are zero off some compact set K which does not depend on m. Therefore, for m large, $\int |f(\boldsymbol{x}) - s_m(\boldsymbol{x})|^p d\mu < \varepsilon^p \mu(K)$ and since ε is arbitrary, this shows that the countable set $\mathscr{D}$ is dense in $L^p(\Omega)$ as claimed. ■

Here is an easier proof if you know the Weierstrass approximation theorem, Theorem 5.6.1 for example.

Theorem 12.3.2 *For $p \geq 1$ and μ a Radon measure, $L^p(\mathbb{R}^n,\mu)$ is separable. Recall this means there exists a countable set $\mathscr{D}$, such that if $f \in L^p(\mathbb{R}^n,\mu)$ and $\varepsilon > 0$, there exists $g \in \mathscr{D}$ such that $\|f-g\|_p < \varepsilon$.*

Proof: As noted above, the continuous functions with compact support are dense in $L^p(\mathbb{R}^n,\mu)$. Let $\mathrm{spt}(f) \subseteq (-R,R)^n$. Consider the polynomials having rational coefficients. By the Weierstrass approximation theorem, and adusting coeffients to make them all rational, there exists p, $\|f-p\|_{[-R,R]^n} < \frac{\varepsilon}{R^{n/p}}$ the norm being the uniform norm $\|g\|_{[-R,R]^n} \equiv \max\{|g(\boldsymbol{x})| : \boldsymbol{x} \in [-R,R]^n\}$. Now let $\mathrm{spt}(f) \prec \tau_\varepsilon \prec V_\varepsilon$ where $\mu(V \setminus \mathrm{spt}(f)) < \varepsilon^p$. Then

$$\begin{aligned} \|f - \tau_\varepsilon p\|_p^p &\leq \int_{\mathrm{spt}(f)} \left(\frac{\varepsilon}{R^{n/p}}\right)^p d\mu + \int_{V\setminus\mathrm{spt}(f)} d\mu \\ &\leq \frac{\varepsilon^p}{R^n} 4^n R^n + \mu(V \setminus \mathrm{spt}(f)) < (1+4^n)\varepsilon^p \end{aligned}$$

Letting $\mathscr{P}$ denote the polynomials with rational coefficients, let $\varepsilon_k \to 0$ and consider the set of functions $\cup_k \tau_{\varepsilon_k}\mathscr{P} \equiv \mathscr{D}$. This is countable and from the above computation, it is dense in L^p. ■

Corollary 12.3.3 *Let Ω be any μ measurable subset of $\mathbb{R}^n$ and let μ be a Radon measure. Then $L^p(\Omega,\mu)$ is separable. Here the σ algebra of measurable sets will consist of all intersections of measurable sets with Ω and the measure will be μ restricted to these sets.*

Proof: Let $\widetilde{\mathscr{D}}$ be the restrictions of $\mathscr{D}$ to Ω. If $f \in L^p(\Omega)$, let F be the zero extension of f to all of $\mathbb{R}^n$. Let $\varepsilon > 0$ be given. By Theorem 12.3.1 or 12.3.2 there exists $s \in \mathscr{D}$ such that $\|F - s\|_p < \varepsilon$. Thus

$$\|s - f\|_{L^p(\Omega,\mu)} \leq \|s - F\|_{L^p(\mathbb{R}^n,\mu)} < \varepsilon$$

and so the countable set $\widetilde{\mathscr{D}}$ is dense in $L^p(\Omega)$. ■

12.4 Continuity of Translation

Definition 12.4.1 *Let f be a function defined on $U \subseteq \mathbb{R}^n$ and let $\boldsymbol{w} \in \mathbb{R}^n$. Then $f_{\boldsymbol{w}}$ will be the function defined on $\boldsymbol{w} + U$ by $f_{\boldsymbol{w}}(\boldsymbol{x}) = f(\boldsymbol{x} - \boldsymbol{w})$.*

Theorem 12.4.2 *(Continuity of translation in L^p) Let $f \in L^p(\mathbb{R}^n)$ with the measure being Lebesgue measure. Then $\lim_{\|\boldsymbol{w}\| \to 0} \|f_{\boldsymbol{w}} - f\|_p = 0$.*

Proof: Let $\varepsilon > 0$ be given and let $g \in C_c(\mathbb{R}^n)$ with $\|g - f\|_p < \frac{\varepsilon}{3}$. Since Lebesgue measure is translation invariant ($m_n(\boldsymbol{w} + E) = m_n(E)$), $\|g_{\boldsymbol{w}} - f_{\boldsymbol{w}}\|_p = \|g - f\|_p < \frac{\varepsilon}{3}$. You can see this from looking at simple functions and passing to the limit or you could use the change of variables formula to verify it.

Therefore

$$\|f - f_{\boldsymbol{w}}\|_p \leq \|f - g\|_p + \|g - g_{\boldsymbol{w}}\|_p + \|g_{\boldsymbol{w}} - f_{\boldsymbol{w}}\| < \frac{2\varepsilon}{3} + \|g - g_{\boldsymbol{w}}\|_p. \tag{12.9}$$

But $\lim_{|\boldsymbol{w}| \to 0} g_{\boldsymbol{w}}(\boldsymbol{x}) = g(\boldsymbol{x})$ uniformly in $\boldsymbol{x}$ because g is uniformly continuous. Now let B be a large ball containing $\operatorname{spt}(g)$ and let δ_1 be small enough that $B(\boldsymbol{x}, \delta) \subseteq B$ whenever $\boldsymbol{x} \in \operatorname{spt}(g)$. If $\varepsilon > 0$ is given there exists $\delta < \delta_1$ such that if $|\boldsymbol{w}| < \delta$, it follows that $|g(\boldsymbol{x} - \boldsymbol{w}) - g(\boldsymbol{x})| < \varepsilon/3 \left(1 + m_n(B)^{1/p}\right)$. Therefore,

$$\|g - g_{\boldsymbol{w}}\|_p = \left(\int_B |g(\boldsymbol{x}) - g(\boldsymbol{x} - \boldsymbol{w})|^p \, dm_n \right)^{1/p} \leq \varepsilon \frac{m_n(B)^{1/p}}{3\left(1 + m_n(B)^{1/p}\right)} < \frac{\varepsilon}{3}.$$

Thus, whenever $|\boldsymbol{w}| < \delta$, it follows $\|g - g_{\boldsymbol{w}}\|_p < \frac{\varepsilon}{3}$ and so from 12.9 $\|f - f_{\boldsymbol{w}}\|_p < \varepsilon$. ■

Here is a remarkable corollary.

Corollary 12.4.3 *Suppose $f \in L^1(\mathbb{R}^p, m_p)$ and let $\boldsymbol{v}$ be any nonzero vector. Then there is a set of measure zero N and a sequence $t_n \to 0+$ such that if $\boldsymbol{x} \notin N$, and $0 < s_n \leq t_n$*

$$\lim_{n \to \infty} |f(\boldsymbol{x}) - f(\boldsymbol{x} + s_n \boldsymbol{v})| = 0.$$

Proof: Let t_n be such that if $s_n \leq t_n$, $\|f - f(\cdot + s_n \boldsymbol{v})\|_{L^1} < 4^{-n}$. This exists by continuity of translation in $L^1(\mathbb{R}^p, m_p)$. Then

$$m_p(E_n) \equiv m_p\left(\left\{\boldsymbol{x} : |f(\boldsymbol{x}) - f(\boldsymbol{x} + s_n \boldsymbol{v})| \geq 2^{-n}\right\}\right) \leq \frac{\int |f - f(\cdot + s_n \boldsymbol{v})| \, dm_p}{2^{-n}} < 2^{-n}$$

Thus there is a set of measure zero N such that if $\boldsymbol{x} \notin N$, then $\boldsymbol{x}$ is in only finitely many of the sets E_n. It follows that for all n sufficiently large $|f(\boldsymbol{x}) - f(\boldsymbol{x} + s_n \boldsymbol{v})| < 2^{-n}$. ■

12.5 Mollifiers and Density of Smooth Functions

Definition 12.5.1 *Let U be an open subset of $\mathbb{R}^n$. $C_c^\infty(U)$ is the vector space of all infinitely differentiable functions which equal zero for all $\boldsymbol{x}$ outside of some compact set contained in U. Similarly, $C_c^m(U)$ is the vector space of all functions which are m times continuously differentiable and whose support is a compact subset of U.*

Example 12.5.2 *Let $U = B(\boldsymbol{z}, 2r)$*

$$\psi(\boldsymbol{x}) = \begin{cases} \exp\left[\left(|\boldsymbol{x}-\boldsymbol{z}|^2 - r^2\right)^{-1}\right] \text{ if } |\boldsymbol{x}-\boldsymbol{z}| < r, \\ 0 \text{ if } |\boldsymbol{x}-\boldsymbol{z}| \geq r. \end{cases}$$

Then a little work shows $\psi \in C_c^\infty(U)$. The following also is easily obtained.

Lemma 12.5.3 *Let U be any open set. Then $C_c^\infty(U) \neq \emptyset$.*

Proof: Pick $\boldsymbol{z} \in U$ and let r be small enough that $B(\boldsymbol{z}, 2r) \subseteq U$. Then let

$$\psi \in C_c^\infty(B(\boldsymbol{z}, 2r)) \subseteq C_c^\infty(U)$$

be the function of the above example.

Definition 12.5.4 *Let $U = \{\boldsymbol{x} \in \mathbb{R}^n : |\boldsymbol{x}| < 1\}$. A sequence $\{\psi_m\} \subseteq C_c^\infty(U)$ is called a mollifier [2] if $\psi_m(\boldsymbol{x}) \geq 0$, $\psi_m(\boldsymbol{x}) = 0$, if $|\boldsymbol{x}| \geq \frac{1}{m}$,and $\int \psi_m(\boldsymbol{x}) = 1$. Sometimes it may be written as $\{\psi_\varepsilon\}$ where ψ_ε satisfies the above conditions except $\psi_\varepsilon(\boldsymbol{x}) = 0$ if $|\boldsymbol{x}| \geq \varepsilon$. In other words, ε takes the place of $1/m$ and in everything that follows $\varepsilon \to 0$ instead of $m \to \infty$.*

As before, $\int f(\boldsymbol{x}, \boldsymbol{y}) d\mu(\boldsymbol{y})$ will mean $\boldsymbol{x}$ is fixed and the function $\boldsymbol{y} \to f(\boldsymbol{x}, \boldsymbol{y})$ is being integrated. To make the notation more familiar, dx is written instead of $dm_n(x)$.

Example 12.5.5 *Let $\psi \in C_c^\infty(B(0,1))$ with $\psi(\boldsymbol{x}) \geq 0$ and $\int \psi dm = 1$. Let $\psi_m(\boldsymbol{x}) = c_m \psi(m\boldsymbol{x})$ where c_m is chosen in such a way that $\int \psi_m dm = 1$. By the change of variables theorem $c_m = m^n$.*

Definition 12.5.6 *A function, f, is said to be in $L^1_{loc}(\mathbb{R}^n, \mu)$ if f is μ measurable and if $|f|\mathcal{X}_K \in L^1(\mathbb{R}^n, \mu)$ for every compact set, K. Here μ is a Radon measure on $\mathbb{R}^n$. Usually $\mu = m_n$, Lebesgue measure. When this is so, write $L^1_{loc}(\mathbb{R}^n)$ or $L^p(\mathbb{R}^n)$, etc. If $f \in L^1_{loc}(\mathbb{R}^n, \mu)$, and $g \in C_c(\mathbb{R}^n)$, $f * g(\boldsymbol{x}) \equiv \int f(\boldsymbol{y}) g(\boldsymbol{x} - \boldsymbol{y}) d\mu$.*

The following lemma will be useful in what follows. It says that one of these very un-regular functions in $L^1_{loc}(\mathbb{R}^n, \mu)$ is smoothed out by convolving with a mollifier.

Lemma 12.5.7 *Let $f \in L^1_{loc}(\mathbb{R}^n, \mu)$, and $g \in C_c^\infty(\mathbb{R}^n)$. Then $f * g$ is an infinitely differentiable function. Here μ is a Radon measure on $\mathbb{R}^n$.*

[2] This is sometimes called an approximate identity if the differentiability is not included.

Proof: Consider the difference quotient for calculating a partial derivative of $f*g$.

$$\frac{f*g(\boldsymbol{x}+t\boldsymbol{e}_j)-f*g(\boldsymbol{x})}{t}=\int f(\boldsymbol{y})\frac{g(\boldsymbol{x}+t\boldsymbol{e}_j-\boldsymbol{y})-g(\boldsymbol{x}-\boldsymbol{y})}{t}d\mu(y).$$

Using the fact that $g\in C_c^\infty(\mathbb{R}^n)$, the quotient $\frac{g(\boldsymbol{x}+t\boldsymbol{e}_j-\boldsymbol{y})-g(\boldsymbol{x}-\boldsymbol{y})}{t}$ is uniformly bounded. To see this easily, use Theorem 7.5.2 on Page 192 to get the existence of a constant, M depending on $\max\{||Dg(\boldsymbol{x})||:\boldsymbol{x}\in\mathbb{R}^n\}$ such that $\left|g(\boldsymbol{x}+t\boldsymbol{e}_j-\boldsymbol{y})-g(\boldsymbol{x}-\boldsymbol{y})\right|\leq M|t|$ for any choice of $\boldsymbol{x}$ and $\boldsymbol{y}$. Therefore, there exists a dominating function for the integrand of the above integral which is of the form $C|f(\boldsymbol{y})|\mathcal{X}_K$ where K is a compact set depending on the support of g. It follows the limit of the difference quotient above passes inside the integral as $t\to 0$ and $\frac{\partial}{\partial x_j}(f*g)(\boldsymbol{x})=\int f(\boldsymbol{y})\frac{\partial}{\partial x_j}g(\boldsymbol{x}-\boldsymbol{y})d\mu(y)$. Now letting $\frac{\partial}{\partial x_j}g$ play the role of g in the above argument, partial derivatives of all orders exist. A similar use of the dominated convergence theorem shows all these partial derivatives are also continuous. ■

Recall also the following partition of unity result. It was proved earlier. See Corollary 12.6.5 on Page 378.

Theorem 12.5.8 *Let K be a compact set in $\mathbb{R}^n$ and let $\{U_i\}_{i=1}^\infty$ be an open cover of K. Then there exist functions, $\psi_k\in C_c^\infty(U_i)$ such that $\psi_i\prec U_i$ and for all $\boldsymbol{x}\in K$, it follows that $\sum_{i=1}^\infty\psi_i(\boldsymbol{x})=1$. If K_1 is a compact subset of U_1 there exist such functions such that also $\psi_1(\boldsymbol{x})=1$ for all $\boldsymbol{x}\in K_1$.*

Note that in the last conclusion of above corollary, the set U_1 could be replaced with U_i for any fixed i by simply renumbering.

Theorem 12.5.9 *For each $p\geq 1$, $C_c^\infty(\mathbb{R}^n)$ is dense in $L^p(\mathbb{R}^n)$. Here the measure is Lebesgue measure.*

Proof: Let $f\in L^p(\mathbb{R}^n)$ and let $\varepsilon>0$ be given. Choose $g\in C_c(\mathbb{R}^n)$ such that $\|f-g\|_p<\frac{\varepsilon}{2}$. This can be done by using Theorem 12.2.4. Now let

$$g_m(\boldsymbol{x})=g*\psi_m(\boldsymbol{x})\equiv\int g(\boldsymbol{x}-\boldsymbol{y})\psi_m(\boldsymbol{y})dm_n(y)=\int g(\boldsymbol{y})\psi_m(\boldsymbol{x}-\boldsymbol{y})dm_n(y)$$

where $\{\psi_m\}$ is a mollifier. It follows from Lemma 12.5.7 $g_m\in C_c^\infty(\mathbb{R}^n)$. It vanishes if $\boldsymbol{x}\notin\operatorname{spt}(g)+B(0,\frac{1}{m})$.

$$\begin{aligned}\|g-g_m\|_p &= \left(\int|g(\boldsymbol{x})-\int g(\boldsymbol{x}-\boldsymbol{y})\psi_m(\boldsymbol{y})dm_n(\boldsymbol{y})|^p dm_n(\boldsymbol{x})\right)^{\frac{1}{p}}\\ &\leq \left(\int(\int|g(\boldsymbol{x})-g(\boldsymbol{x}-\boldsymbol{y})|\psi_m(\boldsymbol{y})dm_n(\boldsymbol{y}))^p dm_n(\boldsymbol{x})\right)^{\frac{1}{p}}\\ &\leq \int\left(\int|g(\boldsymbol{x})-g(\boldsymbol{x}-\boldsymbol{y})|^p dm_n(\boldsymbol{x})\right)^{\frac{1}{p}}\psi_m(\boldsymbol{y})dm_n(\boldsymbol{y})\\ &= \int_{B(0,\frac{1}{m})}\|g-g_{\boldsymbol{y}}\|_p\psi_m(\boldsymbol{y})dm_n(\boldsymbol{y})<\frac{\varepsilon}{2}\end{aligned}$$

whenever m is large enough thanks to the uniform continuity of g. Theorem 12.1.11 was used to obtain the third inequality. There is no measurability problem because the function

$$(\boldsymbol{x},\boldsymbol{y})\to|g(\boldsymbol{x})-g(\boldsymbol{x}-\boldsymbol{y})|\psi_m(\boldsymbol{y})$$

is continuous. Thus when m is large enough,

$$\|f-g_m\|_p \leq \|f-g\|_p + \|g-g_m\|_p < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \blacksquare$$

This is a very remarkable result. Functions in $L^p(\mathbb{R}^n)$ don't need to be continuous anywhere and yet every such function is very close in the L^p norm to one which is infinitely differentiable having compact support. The same result holds for $L^p(U)$ for U an open set. This is the next corollary.

Corollary 12.5.10 *Let U be an open set. For each $p \geq 1$, $C_c^\infty(U)$ is dense in $L^p(U)$. Here the measure is Lebesgue measure.*

Proof: Let $f \in L^p(U)$ and let $\varepsilon > 0$ be given. Choose $g \in C_c(U)$ such that $\|f-g\|_p < \frac{\varepsilon}{2}$. This is possible because Lebesgue measure restricted to the open set, U is regular. Thus the existence of such a g follows from Theorem 12.2.4. Now let

$$g_m(\boldsymbol{x}) = g * \psi_m(\boldsymbol{x}) \equiv \int g(\boldsymbol{x}-\boldsymbol{y})\,\psi_m(\boldsymbol{y})\,dm_n(y) = \int g(\boldsymbol{y})\,\psi_m(\boldsymbol{x}-\boldsymbol{y})\,dm_n(y)$$

where $\{\psi_m\}$ is a mollifier. It follows from Lemma 12.5.7 $g_m \in C_c^\infty(U)$ for all m sufficiently large. It vanishes if $\boldsymbol{x} \notin \mathrm{spt}(g) + B(0, \frac{1}{m})$. Then

$$\begin{aligned}
\|g-g_m\|_p &= \left(\int |g(\boldsymbol{x}) - \int g(\boldsymbol{x}-\boldsymbol{y})\psi_m(\boldsymbol{y})dm_n(\boldsymbol{y})|^p dm_n(\boldsymbol{x})\right)^{\frac{1}{p}} \\
&\leq \left(\int (\int |g(\boldsymbol{x}) - g(\boldsymbol{x}-\boldsymbol{y})|\psi_m(\boldsymbol{y})dm_n(\boldsymbol{y}))^p dm_n(\boldsymbol{x})\right)^{\frac{1}{p}} \\
&\leq \int \left(\int |g(\boldsymbol{x}) - g(\boldsymbol{x}-\boldsymbol{y})|^p dm_n(\boldsymbol{x})\right)^{\frac{1}{p}} \psi_m(\boldsymbol{y})dm_n(\boldsymbol{y}) \\
&= \int_{B(0,\frac{1}{m})} \|g-g_{\boldsymbol{y}}\|_p \psi_m(\boldsymbol{y})dm_n(\boldsymbol{y}) < \frac{\varepsilon}{2}
\end{aligned}$$

whenever m is large enough thanks to uniform continuity of g. Theorem 12.1.11 was used to obtain the third inequality. There is no measurability problem because the function

$$(\boldsymbol{x},\boldsymbol{y}) \to |g(\boldsymbol{x}) - g(\boldsymbol{x}-\boldsymbol{y})|\psi_m(\boldsymbol{y})$$

is continuous. Thus when m is large enough,

$$\|f-g_m\|_p \leq \|f-g\|_p + \|g-g_m\|_p < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \blacksquare$$

Another thing should probably be mentioned. If you have had a course in complex analysis, you may be wondering whether these infinitely differentiable functions having compact support have anything to do with analytic functions which also have infinitely many derivatives. The answer is no! Recall that if an analytic function has a limit point in the set of zeros then it is identically equal to zero. Thus these functions in $C_c^\infty(\mathbb{R}^n)$ are not analytic. This is a strictly real analysis phenomenon and has absolutely nothing to do with the theory of functions of a complex variable.

12.6 Smooth Partitions of Unity

Partitions of unity were discussed earlier. Here the idea of a smooth partition of unity is considered. The earlier general result on metric space is Theorem 3.12.5 on Page 92. Recall the following notation.

Notation 12.6.1 *I will write $\phi \prec V$ to symbolize $\phi \in C_c(V)$, ϕ has values in $[0,1]$, and ϕ has compact support in V. I will write $K \prec \phi \prec V$ for K compact and V open to symbolize ϕ is 1 on K and ϕ has values in $[0,1]$ with compact support contained in V.*

Definition 12.6.2 *A collection of sets $\mathscr{H}$ is called locally finite if for every $\boldsymbol{x}$, there exists $r > 0$ such that $B(\boldsymbol{x}, r)$ has nonempty intersection with only finitely many sets of $\mathscr{H}$. Of course every finite collection of sets is locally finite. This is the case of most interest in this book but the more general notion is interesting.*

The thing about locally finite collection of sets is that the closure of their union equals the union of their closures. This is clearly true of a finite collection.

Lemma 12.6.3 *Let $\mathscr{H}$ be a locally finite collection of sets of a normed vector space V. Then*

$$\overline{\cup \mathscr{H}} = \cup \left\{ \overline{H} : H \in \mathscr{H} \right\}.$$

Proof: It is obvious $\supseteq$ holds in the above claim. It remains to go the other way. Suppose then that $\boldsymbol{p}$ is a limit point of $\cup \mathscr{H}$ and $\boldsymbol{p} \notin \cup \mathscr{H}$. There exists $r > 0$ such that $B(\boldsymbol{p}, r)$ has nonempty intersection with only finitely many sets of $\mathscr{H}$ say these are $H_1, \cdots, H_m$. Then I claim $\boldsymbol{p}$ must be a limit point of one of these. If this is not so, there would exist r' such that $0 < r' < r$ with $B(\boldsymbol{p}, r')$ having empty intersection with each of these H_i. But then $\boldsymbol{p}$ would fail to be a limit point of $\cup \mathscr{H}$. Therefore, $\boldsymbol{p}$ is contained in the right side. It is clear $\cup \mathscr{H}$ is contained in the right side and so This proves the lemma. ■

A good example to consider is the rational numbers each being a set in $\mathbb{R}$. This is **not** a locally finite collection of sets and you note that $\overline{\mathbb{Q}} = \mathbb{R} \neq \cup \{\overline{x} : x \in \mathbb{Q}\}$. By contrast, $\mathbb{Z}$ is a locally finite collection of sets, the sets consisting of individual integers. The closure of $\mathbb{Z}$ is equal to $\mathbb{Z}$ because $\mathbb{Z}$ has no limit points so it contains them all.

Lemma 12.6.4 *Let K be a closed set in $\mathbb{R}^p$ and let $\{V_i\}_{i=1}^{\infty}$ be a locally finite sequence of **bounded** open sets whose union contains K. Then there exist functions, $\psi_i \in C_c^{\infty}(V_i)$ such that for all $\boldsymbol{x} \in K$, $1 = \sum_{i=1}^{\infty} \psi_i(\boldsymbol{x})$ and the function $f(\boldsymbol{x})$ given by $f(\boldsymbol{x}) = \sum_{i=1}^{\infty} \psi_i(\boldsymbol{x})$ is in $C^{\infty}(\mathbb{R}^p)$.*

Proof: Let $K_1 = K \setminus \cup_{i=2}^{\infty} V_i$. Thus K_1 is compact because it is a closed subset of a bounded set and $K_1 \subseteq V_1$. Let W_1 be an open set having compact closure which satisfies

$$K_1 \subseteq W_1 \subseteq \overline{W}_1 \subseteq V_1$$

Thus $W_1, V_2, \cdots$ covers K and $\overline{W}_1 \subseteq V_1$. Suppose $W_1, \cdots, W_r$ have been defined such that $\overline{W_i} \subseteq V_i$ for each i, and $W_1, \cdots, W_r, V_{r+1}, \cdots$ covers K. Then let

$$K_{r+1} \equiv K \setminus \left(\left(\cup_{i=r+2}^{\infty} V_i \right) \cup \left(\cup_{j=1}^{r} W_j \right) \right).$$

It follows K_{r+1} is compact because $K_{r+1} \subseteq V_{r+1}$. Let W_{r+1} satisfy

$$K_{r+1} \subseteq W_{r+1} \subseteq \overline{W}_{r+1} \subseteq V_{r+1}, \ \overline{W}_{r+1} \text{ is compact}$$

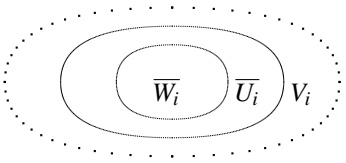
Continuing this way defines a sequence of open sets $\{W_i\}_{i=1}^{\infty}$ having compact closures with the property

$$\overline{W_i} \subseteq V_i,\ K \subseteq \cup_{i=1}^{\infty} W_i.$$

Note $\{W_i\}_{i=1}^{\infty}$ is locally finite because the original sequence, $\{V_i\}_{i=1}^{\infty}$ was locally finite. Now let U_i be open sets which satisfy

$$\overline{W}_i \subseteq U_i \subseteq \overline{U}_i \subseteq V_i,\ \overline{U}_i \text{ is compact.}$$

Similarly, $\{U_i\}_{i=1}^{\infty}$ is locally finite.



Now the local finiteness implies $\overline{\cup_{i=1}^{\infty} W_i} = \cup_{i=1}^{\infty} \overline{W_i}$ and so it is possible from Theorem 12.5.8 to define ϕ_i and γ, infinitely differentiable functions having compact support such that

$$\overline{U}_i \prec \phi_i \prec V_i,\ \cup_{i=1}^{\infty} \overline{W}_i \prec \gamma \prec \cup_{i=1}^{\infty} U_i.$$

Now define

$$\psi_i(\boldsymbol{x}) = \begin{cases} \gamma(\boldsymbol{x})\phi_i(\boldsymbol{x}) / \sum_{j=1}^{\infty} \phi_j(\boldsymbol{x}) \text{ if } \sum_{j=1}^{\infty} \phi_j(\boldsymbol{x}) \neq 0, \\ 0 \text{ if } \sum_{j=1}^{\infty} \phi_j(\boldsymbol{x}) = 0. \end{cases}$$

All of these infinite sums are really finite sums because of the local finiteness of the $\{V_i\}$. Thus for $\boldsymbol{y}$ near a given $\boldsymbol{x}$, all $\phi_j(\boldsymbol{y})$ are zero. Therefore, all continuity and differentiability of the individual ϕ_j is retained by the "infinite" sum.

If $\boldsymbol{x}$ is such that $\sum_{j=1}^{\infty} \phi_j(\boldsymbol{x}) = 0$, then $\boldsymbol{x} \notin \cup_{i=1}^{\infty} \overline{U_i}$ because ϕ_i equals one on $\overline{U_i}$. Consequently $\gamma(\boldsymbol{y}) = 0$ for all $\boldsymbol{y}$ near $\boldsymbol{x}$ thanks to the fact that $\cup_{i=1}^{\infty} \overline{U_i}$ is closed and so $\psi_i(\boldsymbol{y}) = 0$ for all $\boldsymbol{y}$ near $\boldsymbol{x}$. Hence ψ_i is infinitely differentiable at such $\boldsymbol{x}$. If $\sum_{j=1}^{\infty} \phi_j(\boldsymbol{x}) \neq 0$, this situation persists near $\boldsymbol{x}$ because each ϕ_j is continuous and so ψ_i is infinitely differentiable at such points also. Therefore ψ_i is infinitely differentiable. If $\boldsymbol{x} \in K$, then $\gamma(\boldsymbol{x}) = 1$ and so $\sum_{j=1}^{\infty} \psi_j(\boldsymbol{x}) = 1$. Clearly $0 \leq \psi_i(\boldsymbol{x}) \leq 1$ and $\text{spt}(\psi_j) \subseteq V_j$. ■

The functions, $\{\psi_i\}$ are called a C^{∞} partition of unity. The following is very useful.

Corollary 12.6.5 *If H is a compact subset of V_i for some V_i there exists a partition of unity such that $\psi_i(x) = 1$ for all $x \in H$ in addition to the conclusion of Lemma 12.6.4.*

Proof: Keep V_i the same but replace all the V_j with $\widetilde{V}_j \equiv V_j \setminus H$. Now in the proof above, applied to this modified collection of open sets, if $j \neq i, \phi_j(x) = 0$ whenever $x \in H$. Therefore, $\psi_i(x) = 1$ on H. ■

12.7 Exercises

1. Let E be a Lebesgue measurable set in $\mathbb{R}$. Suppose $m(E) > 0$. Consider the set $E - E = \{x - y : x \in E, y \in E\}$. Show that $E - E$ contains an interval. **Hint:** Let $f(x) = \int \mathcal{X}_E(t)\mathcal{X}_E(x+t)dt$. Note f is continuous at 0 and $f(0) > 0$ and use continuity of translation in L^p.

2. Establish the inequality $\|fg\|_r \leq \|f\|_p \|g\|_q$ whenever $\frac{1}{r} = \frac{1}{p} + \frac{1}{q}$.

3. Let $(\Omega, \mathscr{S}, \mu)$ be counting measure on $\mathbb{N}$. Thus $\Omega = \mathbb{N}$ and $\mathscr{S} = \mathscr{P}(\mathbb{N})$ with $\mu(S) =$ number of things in S. Let $1 \leq p \leq q$. Show that in this case, $L^1(\mathbb{N}) \subseteq L^p(\mathbb{N}) \subseteq L^q(\mathbb{N})$. **Hint:** This is real easy if you consider what $\int_\Omega f d\mu$ equals. How are the norms related?

4. Consider the function, $f(x,y) = \frac{x^{p-1}}{py} + \frac{y^{q-1}}{qx}$ for $x, y > 0$ and $\frac{1}{p} + \frac{1}{q} = 1$. Show directly that $f(x,y) \geq 1$ for all such x, y and show this implies $xy \leq \frac{x^p}{p} + \frac{y^q}{q}$.

5. Give an example of a sequence of functions in $L^p(\mathbb{R})$ which converges to zero in L^p but does not converge pointwise to 0. Does this contradict the proof of the theorem that L^p is complete?

6. Let K be a bounded subset of $L^p(\mathbb{R}^n)$ and suppose that there exists G such that $\overline{G}$ is compact with $\int_{\mathbb{R}^n \setminus \overline{G}} |u(\boldsymbol{x})|^p \, dx < \varepsilon^p$ and for all $\varepsilon > 0$, there exist a $\delta > 0$ and such that if $|\boldsymbol{h}| < \delta$, then $\int |u(\boldsymbol{x} + \boldsymbol{h}) - u(\boldsymbol{x})|^p \, dx < \varepsilon^p$ for all $u \in K$. Show that K is precompact in $L^p(\mathbb{R}^n)$. **Hint:** Let ϕ_k be a mollifier and consider $K_k \equiv \{u * \phi_k : u \in K\}$. Verify the conditions of the Ascoli Arzela theorem for these functions defined on $\overline{G}$ and show there is an ε net for each $\varepsilon > 0$. Can you modify this to let an arbitrary open set take the place of $\mathbb{R}^n$?

7. Let (Ω, d) be a metric space and suppose also that $(\Omega, \mathscr{S}, \mu)$ is a regular measure space such that $\mu(\Omega) < \infty$ and let $f \in L^1(\Omega)$ where f has complex values. Show that for every $\varepsilon > 0$, there exists an open set of measure less than ε, denoted here by V and a continuous function, g defined on Ω such that $f = g$ on V^C. Thus, aside from a set of small measure, f is continuous. If $|f(\omega)| \leq M$, show that it can be assumed that $|g(\omega)| \leq M$. This is called Lusin's theorem. **Hint:** Use Theorems 12.2.4 and 12.1.9 to obtain a sequence of functions in $C_c(\Omega), \{g_n\}$ which converges pointwise a.e. to f and then use Egoroff's theorem to obtain a small set, W of measure less than $\varepsilon/2$ such that convergence is uniform on W^C. Now let F be a closed subset of W^C such that $\mu(W^C \setminus F) < \varepsilon/2$. Let $V = F^C$. Thus $\mu(V) < \varepsilon$ and on $F = V^C$, the convergence of $\{g_n\}$ is uniform showing that the restriction of f to V^C is continuous. Now use the Tietze extension theorem.

8. Let $\phi_m \in C_c^\infty(\mathbb{R}^n), \phi_m(\boldsymbol{x}) \geq 0$ and $\int_{\mathbb{R}^n} \phi_m(\boldsymbol{y}) dy = 1$ with
$$\lim_{m \to \infty} \sup \{|\boldsymbol{x}| : \boldsymbol{x} \in \operatorname{spt}(\phi_m)\} = 0.$$
Show if $f \in L^p(\mathbb{R}^n)$, $\lim_{m \to \infty} f * \phi_m = f$ in $L^p(\mathbb{R}^n)$.

9. Let $\phi : \mathbb{R} \to \mathbb{R}$ be convex. This means $\phi(\lambda x + (1-\lambda)y) \leq \lambda \phi(x) + (1-\lambda)\phi(y)$ whenever $\lambda \in [0,1]$. Verify that if $x < y < z$, then $\frac{\phi(y) - \phi(x)}{y - x} \leq \frac{\phi(z) - \phi(y)}{z - y}$ and that $\frac{\phi(z) - \phi(x)}{z - x} \leq \frac{\phi(z) - \phi(y)}{z - y}$. Show if $s \in \mathbb{R}$ there exists λ such that $\phi(s) \leq \phi(t) + \lambda(s - t)$ for all t. Show that if ϕ is convex, then ϕ is continuous.

10. $\uparrow$ Prove Jensen's inequality. If $\phi : \mathbb{R} \to \mathbb{R}$ is convex, $\mu(\Omega) = 1$, and $f : \Omega \to \mathbb{R}$ is in $L^1(\Omega)$, then $\phi(\int_\Omega f \, du) \leq \int_\Omega \phi(f) d\mu$. **Hint:** Let $s = \int_\Omega f \, d\mu$ and use Problem 9.

11. Let $\frac{1}{p} + \frac{1}{p'} = 1$, $p > 1$, let $f \in L^p(\mathbb{R})$, $g \in L^{p'}(\mathbb{R})$. Show $f * g$ is uniformly continuous on $\mathbb{R}$ and $|(f * g)(x)| \leq \|f\|_{L^p} \|g\|_{L^{p'}}$. **Hint:** You need to consider why $f * g$ exists and then this follows from the definition of convolution and continuity of translation in L^p.

12. $B(p,q) = \int_0^1 x^{p-1}(1-x)^{q-1}dx, \Gamma(p) = \int_0^\infty e^{-t}t^{p-1}dt$ for $p,q>0$. The first of these is called the beta function, while the second is the gamma function. Show a.) $\Gamma(p+1) = p\Gamma(p)$; b.) $\Gamma(p)\Gamma(q) = B(p,q)\Gamma(p+q)$.

13. Let $f \in C_c(0,\infty)$. Define $F(x) = \frac{1}{x}\int_0^x f(t)dt$. Show $\|F\|_{L^p(0,\infty)} \leq \frac{p}{p-1}\|f\|_{L^p(0,\infty)}$ whenever $p>1$. **Hint:** Argue there is no loss of generality in assuming $f \geq 0$ and then assume this is so. Integrate $\int_0^\infty |F(x)|^p dx$ by parts as follows:

$$\int_0^\infty F^p dx = \overbrace{xF^p|_0^\infty}^{\text{show = 0}} - p\int_0^\infty xF^{p-1}F'dx.$$

Now show $xF' = f - F$ and use this in the last integral. Complete the argument by using Holder's inequality and $p-1 = p/q$.

14. $\uparrow$ Now suppose $f \in L^p(0,\infty)$, $p>1$, and f not necessarily in $C_c(0,\infty)$. Show that $F(x) = \frac{1}{x}\int_0^x f(t)dt$ still makes sense for each $x>0$. Show the inequality of Problem 13 is still valid. This inequality is called Hardy's inequality. **Hint:** To show this, use the above inequality along with the density of $C_c(0,\infty)$ in $L^p(0,\infty)$.

15. Suppose $f,g \geq 0$. When does equality hold in Holder's inequality?

16. Show the Vitali Convergence theorem implies the Dominated Convergence theorem for finite measure spaces but there exist examples where the Vitali convergence theorem works and the dominated convergence theorem does not.

17. $\uparrow$ Suppose $\mu(\Omega) < \infty$, $\{f_n\} \subseteq L^1(\Omega)$, and $\int_\Omega h(|f_n|)\,d\mu < C$ for all n where h is a continuous, nonnegative function satisfying $\lim_{t\to\infty}\frac{h(t)}{t} = \infty$. Show $\{f_n\}$ is uniformly integrable. In applications, this often occurs in the form of a bound on $\|f_n\|_p$.

18. $\uparrow$ Sometimes, especially in books on probability, a different definition of uniform integrability is used than that presented here. A set of functions $\mathfrak{S}$, defined on a finite measure space, $(\Omega,\mathscr{S},\mu)$ is said to be uniformly integrable if for all $\varepsilon>0$ there exists $\alpha>0$ such that for all $f \in \mathfrak{S}$, $\int_{[|f|\geq\alpha]}|f|\,d\mu \leq \varepsilon$. Show that this definition is equivalent to the definition of uniform integrability with the addition of the condition that there is a constant, $C<\infty$ such that $\int |f|\,d\mu \leq C$ for all $f \in \mathfrak{S}$.

19. Suppose $f \in L^\infty \cap L^1$. Show $\lim_{p\to\infty}\|f\|_{L^p} = \|f\|_\infty$. **Hint:**

$$(||f||_\infty - \varepsilon)^p \mu([|f| > ||f||_\infty - \varepsilon]) \leq \int_{[|f|>||f||_\infty-\varepsilon]} |f|^p\,d\mu \leq$$

$$\int |f|^p\,d\mu = \int |f|^{p-1}|f|\,d\mu \leq ||f||_\infty^{p-1}\int |f|\,d\mu.$$

Now raise both ends to the $1/p$ power and take liminf and limsup as $p\to\infty$. You should get $||f||_\infty - \varepsilon \leq \liminf ||f||_p \leq \limsup ||f||_p \leq ||f||_\infty$

20. Suppose $\mu(\Omega)<\infty$. Show that if $1 \leq p < q$, then $L^q(\Omega) \subseteq L^p(\Omega)$. **Hint** Use Holder's inequality.

21. Show $L^1(\mathbb{R}) \not\subseteq L^2(\mathbb{R})$ and $L^2(\mathbb{R}) \not\subseteq L^1(\mathbb{R})$ if Lebesgue measure is used. **Hint:** Consider $1/\sqrt{x}$ and $1/x$.

22. Suppose that $\theta \in [0,1]$ and $r,s,q>0$ with $\frac{1}{q}=\frac{\theta}{r}+\frac{1-\theta}{s}$. Show that $\left(\int |f|^q d\mu\right)^{1/q} \leq \left(\int |f|^r d\mu\right)^{1/r)^{\theta}} \left(\int |f|^s d\mu\right)^{1/s)^{1-\theta}}$. If $q,r,s \geq 1$ this says that $\|f\|_q \leq \|f\|_r^{\theta} \|f\|_s^{1-\theta}$. Using this, show that $\ln\left(\|f\|_q\right) \leq \theta \ln\left(\|f\|_r\right) + (1-\theta)\ln\left(\|f\|_s\right)$. **Hint:** $\int |f|^q d\mu = \int |f|^{q\theta} |f|^{q(1-\theta)} d\mu$.Now note that $1=\frac{\theta q}{r}+\frac{q(1-\theta)}{s}$ and use Holder's inequality.

23. Suppose f is a function in $L^1(\mathbb{R})$ and f is infinitely differentiable. Is $f' \in L^1(\mathbb{R})$? **Hint:** What if $\phi \in C_c^{\infty}(0,1)$ and $f(x)=\phi\left(2^n(x-n)\right)$ for $x \in (n,n+1)$, $f(x)=0$ if $x<0$?

24. Establish the following for $f \neq 0$ in L^p

$$\|f\|_p = \int_{\Omega} \frac{|f|^p}{\|f\|_p^{p-1}} d\mu = \int_{\Omega} f \frac{|f|^{p-2}\bar{f}}{\|f\|_p^{p-1}} \leq \sup_{\|g\|_q \leq 1} \int |f|\,|g|\,d\mu \leq \|f\|_p$$

25. ↑From the above problem, if f is nonnegative and product measurable,

$$\left(\int \left(\int f(x,y)\,d\mu(x)\right)^p d\nu(y)\right)^{1/p} = \sup_{\|h\|_q \leq 1} \int \left(\int f(x,y)\,d\mu(x)\right) h(y)\,d\nu(y)$$

Now use Fubini's theorem and then the Holder inequality to obtain

$$= \sup_{\|h\|_q \leq 1} \int \int f(x,y)\,h(y)\,d\nu(y)\,d\mu(x) \leq \int \left(\int f(x,y)^p\,d\nu(y)\right)^{1/p} d\mu(x)$$

This gives another proof of the important Minkowski inequality for integrals.

26. Let $0<p<1$ and let f,g be measurable $\mathbb{C}$ valued functions. Also

$$\int_{\Omega} |g|^{p/(p-1)}\,d\mu < \infty,\ \int_{\Omega} |f|^p\,d\mu < \infty$$

Then show the following backwards Holder inequality holds.

$$\int_{\Omega} |fg|\,d\mu \geq \left(\int_{\Omega} |f|^p\,d\mu\right)^{1/p} \left(\int_{\Omega} |g|^{p/(p-1)}\,d\mu\right)^{(p-1)/p}$$

Hint: You should first note that $g=0$ only on a set of measure zero. Then you could write $\int_{\Omega} |f|^p\,d\mu = \int_{\Omega} |g|^{-p} |fg|^p\,d\mu$. Apply the usual Holder inequality with $1/p$ one of the exponents and $1/(1-p)$ the other exponent. Then the above is $\leq \left(\int_{\Omega} |g|^{-p/(1-p)}\,d\mu\right)^{1-p} \left(\int_{\Omega} |fg|\,d\mu\right)^p$ etc. Note the usual Holder inequality in case $p>1$ is $\int_{\Omega} |fg|\,d\mu \leq \left(\int_{\Omega} |f|^p\,d\mu\right)^{1/p} \left(\int_{\Omega} |g|^{p/(p-1)}\,d\mu\right)^{(p-1)/p}$.

27. Let $0<p<1$ and suppose $\int |h|^p\,d\mu < \infty$ for $h=f,g$. Then

$$\left(\int (|f|+|g|)^p\,d\mu\right)^{1/p} \geq \left(\int |f|^p\,d\mu\right)^{1/p} + \left(\int |g|^p\,d\mu\right)^{1/p}$$

This is the backwards Minkowski inequality. **Hint:** First explain why, since $p < 1, (|f|+|g|)^p \leq |f|^p + |g|^p$. It follows from this that

$$\int_\Omega \left((|f|+|g|)^{p-1} \right)^{p/(p-1)} d\mu < \infty$$

since $\int |h|^p \, d\mu < \infty$ for $h = f, g$. Then $(|f|+|g|)^{p-1}$ plays the role of $|g|$ in the above backwards Holder inequality. Next do this:

$$\int_\Omega (|f|+|g|)^p \, d\mu = \int_\Omega (|f|+|g|)^{p-1} (|f|+|g|) \, d\mu$$

$$= \int_\Omega (|f|+|g|)^{p-1} |f| \, d\mu + \int_\Omega (|f|+|g|)^{p-1} |g| \, d\mu$$

Now apply the backwards Holder inequality of the above problem. The first term gives

$$\int_\Omega (|f|+|g|)^{p-1} |f| \, d\mu \geq \left(\int_\Omega (|f|+|g|)^p \, d\mu \right)^{(p-1)/p} \left(\int_\Omega |f|^p \, d\mu \right)^{1/p}$$

28. Let $f \in L^1_{loc}(\mathbb{R})$. Show there exists a set of measure zero N such that if $x \notin N$, then if $\{I_n\}$ is a sequence of intervals containing x such that $m(I_n) \to 0$ then

$$\frac{1}{m(I_n)} \int_{I_n} |f - f(x)| \, dx \to 0.$$

Generalize to higher dimensions if possible. Also, does I_n have to be an interval?

29. Suppose $F(x) = \int_a^x f(t) \, dt$ so that F is absolutely continuous where $f \in L^1([a,b])$. Show that $f \in L^p$ for $p > 1$ if and only if there is $M < \infty$ such that whenever $a = x_0 < x_1 < \cdots < x_n = b$ it follows that $\sum_{i=1}^n \frac{|F(x_i) - F(x_{i-1})|^p}{(x_i - x_{i-1})^{p-1}} < M$. This problem is a result of F. Riesz. **Hint:** The first part is an easy application of Holder's inequality. For the second, let $\mathscr{P}_n$ be a sequence of paritions of $[a,b]$ such that the subintervals have lengths converging to 0. Define $f_n(x) \equiv \sum_{k=1}^n \frac{F(x_k^n) - F(x_{k-1}^n)}{x_k^n - x_{k-1}^n} \mathscr{X}_{I_k^n}(x)$ where the intervals of $\mathscr{P}_n$ are $I_k^n = [x_{k-1}^n, x_{k+1}^n]$. Then for a.e. $x, f_n(x) \to f(x)$ thanks to the Lebesgue fundamental theorem of calculus and Problem 28. Now apply Fatou's lemma to say that $\int_a^b |f(x)|^p \, dx \leq \liminf_{n\to\infty} \int_a^b |f_n(x)|^p \, dx$ and simplify this last integral by breaking it into a sum of integrals over the sub-intervals of $\mathscr{P}_n$. Note $\frac{|F(x_k^n) - F(x_{k-1}^n)|^p}{(x_k^n - x_{k-1}^n)^p}$ does not depend on $x \in I_k^n$.

30. If $f \in L^p(U, m_p)$, where U is a bounded open set in $\mathbb{R}^n$, show there exists a sequence of smooth functions which converges to f a.e. Then show there exists a sequence of polynomials whose restriction to U converges a.e. to f on U.

31. In Corollary 12.4.3 can you generalize where f is only in $L^1_{loc}(\mathbb{R}^p, m_p)$.

Chapter 13

Fourier Transforms

13.1 Fourier Transforms of Functions in $\mathscr{G}$

First is a definition of a very specialized set of functions. Here the measure space will be $(\mathbb{R}^n, m_n, \mathscr{F}_n)$, familiar Lebesgue measure.

First recall the following definition of a polynomial.

Definition 13.1.1 *$\alpha = (\alpha_1, \cdots, \alpha_n)$ for $\alpha_1 \cdots \alpha_n$ nonnegative integers is called a multi-index. For α a multi-index, $|\alpha| \equiv \alpha_1 + \cdots + \alpha_n$ and if $\boldsymbol{x} \in \mathbb{R}^n$,$\boldsymbol{x} = (x_1, \cdots, x_n)$, and f a function, define*

$$\boldsymbol{x}^\alpha \equiv x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}.$$

A polynomial in n variables of degree m is a function of the form

$$p(\boldsymbol{x}) = \sum_{|\alpha| \leq m} a_\alpha \boldsymbol{x}^\alpha.$$

Here α is a multi-index as just described and $a_\alpha \in \mathbb{C}$. Also define for $\alpha = (\alpha_1, \cdots, \alpha_n)$ a multi-index

$$D^\alpha f(\boldsymbol{x}) \equiv \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \cdots \partial x_n^{\alpha_n}}.$$

Definition 13.1.2 *Define $\mathscr{G}_1$ to be the functions of the form $p(\boldsymbol{x}) e^{-a|\boldsymbol{x}|^2}$ where $a > 0$ is rational and $p(\boldsymbol{x})$ is a polynomial having all rational coefficients, a_α being "rational" if it is of the form $a + ib$ for $a, b \in \mathbb{Q}$. Let $\mathscr{G}$ be all finite sums of functions in $\mathscr{G}_1$. Thus $\mathscr{G}$ is an algebra of functions which has the property that if $f \in \mathscr{G}$ then $\overline{f} \in \mathscr{G}$.*

Thus there are countably many functions in $\mathscr{G}_1$. This is because, for each m, there are countably many choices for a_α for $|\alpha| \leq m$ since there are finitely many α for $|\alpha| \leq m$ and for each such α, there are countably many choices for a_α since $\mathbb{Q}+i\mathbb{Q}$ is countable. (Why?) Thus there are countably many polynomials having degree no more than m. This is true for each m and so the number of different polynomials is a countable union of countable sets which is countable. Now there are countably many choices of $e^{-\alpha|\boldsymbol{x}|^2}$ and so there are countably many in $\mathscr{G}_1$ because the Cartesian product of countable sets is countable.

Now $\mathscr{G}$ consists of finite sums of functions in $\mathscr{G}_1$. Therefore, it is countable because for each $m \in \mathbb{N}$, there are countably many such sums which are possible.

I will show now that $\mathscr{G}$ is dense in $L^p(\mathbb{R}^n)$ but first, here is a lemma which follows from the Stone Weierstrass theorem.

Lemma 13.1.3 *$\mathscr{G}$ is dense in $C_0(\mathbb{R}^n)$ with respect to the norm,*

$$||f||_\infty \equiv \sup\{|f(\boldsymbol{x})| : \boldsymbol{x} \in \mathbb{R}^n\}$$

Proof: By the Weierstrass approximation theorem, it suffices to show $\mathscr{G}$ separates the points and annihilates no point. It was already observed in the above definition that $\overline{f} \in \mathscr{G}$ whenever $f \in \mathscr{G}$. If $\boldsymbol{y}_1 \neq \boldsymbol{y}_2$ suppose first that $|\boldsymbol{y}_1| \neq |\boldsymbol{y}_2|$. Then in this case, you can let $f(\boldsymbol{x}) \equiv e^{-|\boldsymbol{x}|^2}$. Then $f \in \mathscr{G}$ and $f(\boldsymbol{y}_1) \neq f(\boldsymbol{y}_2)$. If $|\boldsymbol{y}_1| = |\boldsymbol{y}_2|$, then suppose $y_{1k} \neq y_{2k}$. This must happen for some k because $\boldsymbol{y}_1 \neq \boldsymbol{y}_2$. Then let $f(\boldsymbol{x}) \equiv x_k e^{-|\boldsymbol{x}|^2}$. Thus $\mathscr{G}$ separates points. Now $e^{-|\boldsymbol{x}|^2}$ is never equal to zero and so $\mathscr{G}$ annihilates no point of $\mathbb{R}^n$. ■

These functions are clearly quite specialized. Therefore, the following theorem is somewhat surprising.

Theorem 13.1.4 *For each $p \geq 1, p < \infty, \mathscr{G}$ is dense in $L^p(\mathbb{R}^n)$. Since $\mathscr{G}$ is countable, this shows that $L^p(\mathbb{R}^n)$ is separable.*

Proof: Let $f \in L^p(\mathbb{R}^n)$. Then there exists $g \in C_c(\mathbb{R}^n)$ such that $\|f-g\|_p < \varepsilon$. Now let $b > 0$ be large enough that

$$\int_{\mathbb{R}^n} \left(e^{-b|\boldsymbol{x}|^2}\right)^p dx < \varepsilon^p.$$

Then $\boldsymbol{x} \to g(\boldsymbol{x}) e^{b|\boldsymbol{x}|^2}$ is in $C_c(\mathbb{R}^n) \subseteq C_0(\mathbb{R}^n)$. Therefore, from Lemma 13.1.3 there exists $\psi \in \mathscr{G}$ such that

$$\left\| g e^{b|\cdot|^2} - \psi \right\|_\infty < 1$$

Therefore, letting $\phi(\boldsymbol{x}) \equiv e^{-b|\boldsymbol{x}|^2} \psi(\boldsymbol{x})$ it follows that $\phi \in \mathscr{G}$ and for all $\boldsymbol{x} \in \mathbb{R}^n$,

$$|g(\boldsymbol{x}) - \phi(\boldsymbol{x})| < e^{-b|\boldsymbol{x}|^2}$$

Therefore, $\left(\int_{\mathbb{R}^n} |g(\boldsymbol{x}) - \phi(\boldsymbol{x})|^p dx\right)^{1/p} \leq \left(\int_{\mathbb{R}^n} \left(e^{-b|\boldsymbol{x}|^2}\right)^p dx\right)^{1/p} < \varepsilon$. It follows

$$\|f - \phi\|_p \leq \|f-g\|_p + \|g-\phi\|_p < 2\varepsilon. \blacksquare$$

From now on, we can drop the restriction that the coefficients of the polynomials in $\mathscr{G}$ are rational. We also drop the restriction that a is rational. Thus $\mathscr{G}$ will be finite sums of functions which are of the form $p(\boldsymbol{x}) e^{-a|\boldsymbol{x}|^2}$ where the coefficients of p are complex and $a > 0$.

The following lemma is also interesting even if it is obvious.

Lemma 13.1.5 *For $\psi \in \mathscr{G}$, p a polynomial, and α, β multi-indices, $D^\alpha \psi \in \mathscr{G}$ and $p\psi \in \mathscr{G}$. Also*

$$\sup\{|\boldsymbol{x}^\beta D^\alpha \psi(\boldsymbol{x})| : \boldsymbol{x} \in \mathbb{R}^n\} < \infty$$

Thus these special functions are infinitely differentiable (smooth). They also have the property that they and all their partial derivatives vanish as $|\boldsymbol{x}| \to \infty$.

Let $\mathscr{G}$ be the functions of Definition 13.1.2 except, for the sake of convenience, remove all references to rational numbers. Thus $\mathscr{G}$ consists of finite sums of polynomials having coefficients in $\mathbb{C}$ times $e^{-a|\boldsymbol{x}|^2}$ for some $a > 0$. The idea is to first understand the Fourier transform on these very specialized functions.

Definition 13.1.6 *For $\psi \in \mathscr{G}$ Define the Fourier transform F and the inverse Fourier transform, F^{-1} by*

$$F\psi(\boldsymbol{t}) \equiv (2\pi)^{-n/2} \int_{\mathbb{R}^n} e^{-i\boldsymbol{t}\cdot\boldsymbol{x}} \psi(\boldsymbol{x}) dx, \quad F^{-1}\psi(\boldsymbol{t}) \equiv (2\pi)^{-n/2} \int_{\mathbb{R}^n} e^{i\boldsymbol{t}\cdot\boldsymbol{x}} \psi(\boldsymbol{x}) dx.$$

where $\boldsymbol{t}\cdot\boldsymbol{x} \equiv \sum_{i=1}^n t_i x_i$. Note there is no problem with this definition because ψ is in $L^1(\mathbb{R}^n)$ and therefore, $\left|e^{i\boldsymbol{t}\cdot\boldsymbol{x}} \psi(\boldsymbol{x})\right| \leq |\psi(\boldsymbol{x})|$, an integrable function.

The following lemma follows from the dominated convergence theorem and routine manipulations.

Lemma 13.1.7 *For $\psi \in \mathcal{G}$,*

$$D_t^\alpha F(\psi) = (-i)^{|\alpha|} F(x^\alpha \psi(x)), \quad D_t^\alpha F^{-1}(\psi) = (i)^{|\alpha|} F^{-1}(x^\alpha \psi(x))$$

In this lemma, $x^\alpha \psi(x)$ denotes the function $x \to x^\alpha \psi(x)$.

One reason for using the functions $\mathcal{G}$ is that it is very easy to compute the Fourier transform of these functions. The first thing to do is to verify F and F^{-1} map $\mathcal{G}$ to $\mathcal{G}$ and that $F^{-1} \circ F(\psi) = \psi$. Next is a simple lemma from calculus which is about the Fourier transforms of $e^{-c|t|^2}$.

Lemma 13.1.8 *The following hold.* $(c > 0)$

$$\left(\frac{1}{2\pi}\right)^{n/2} \int_{\mathbb{R}^n} e^{-c|t|^2} e^{-is\cdot t} dt = \left(\frac{1}{2\pi}\right)^{n/2} \int_{\mathbb{R}^n} e^{-c|t|^2} e^{is\cdot t} dt$$

$$= \left(\frac{1}{2\pi}\right)^{n/2} e^{-\frac{|s|^2}{4c}} \left(\frac{\sqrt{\pi}}{\sqrt{c}}\right)^n = \left(\frac{1}{2c}\right)^{n/2} e^{-\frac{1}{4c}|s|^2}. \tag{13.1}$$

That is, $F\left(e^{-c|t|^2}\right) = \left(\frac{1}{2c}\right)^{n/2} e^{-\frac{1}{4c}|s|^2}$ *and* $F^{-1}\left(e^{-c|t|^2}\right) = \left(\frac{1}{2c}\right)^{n/2} e^{-\frac{1}{4c}|s|^2}$.

Proof: Consider first the case of one dimension. Let $H(s)$ be given by

$$H(s) \equiv \int_{\mathbb{R}} e^{-ct^2} e^{-ist} dt = \int_{\mathbb{R}} e^{-ct^2} \cos(st)\, dt$$

Then $H(0)^2 = \int_{\mathbb{R}} \int_{\mathbb{R}} e^{-c\left(t^2+s^2\right)} dt ds = \int_0^\infty \int_0^{2\pi} e^{-cr^2} r d\theta dr = \frac{1}{c}\pi$ so $H(0) = \sqrt{\frac{\pi}{c}}$. Then using the dominated convergence theorem to differentiate, $H'(s) + \frac{s}{2c} H(s) = 0, H(0) = \sqrt{\frac{\pi}{c}}$. Thus

$$\frac{d}{ds}\left(e^{s^2/4c} H(s)\right) = 0 \text{ and so } e^{s^2/4c} H(s) - \sqrt{\frac{\pi}{c}} = 0, \text{ so } H(s) = \sqrt{\frac{\pi}{c}} e^{-s^2/4c}$$

Hence

$$\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-ct^2} e^{-ist} dt = \sqrt{\frac{\pi}{c}} \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{4c}} = \left(\frac{1}{2c}\right)^{1/2} e^{-\frac{s^2}{4c}}.$$

This proves the formula in the case of one dimension. The case of the inverse Fourier transform is similar. The n dimensional formula follows from Fubini's theorem. ■

With these formulas, it is easy to verify F, F^{-1} map $\mathcal{G}$ to $\mathcal{G}$ and $F \circ F^{-1} = F^{-1} \circ F = id$.

Theorem 13.1.9 *Each of F and F^{-1} map $\mathcal{G}$ to $\mathcal{G}$. Also $F^{-1} \circ F(\psi) = \psi$ and $F \circ F^{-1}(\psi) = \psi$.*

Proof: It is obvious that F, F^{-1} map $\mathcal{G}$ to $\mathcal{G}$ using Lemmas 13.1.7, 13.1.8. Indeed, for $\psi(x) = x^\alpha e^{-c|x|^2}$,

$$\begin{aligned} F(\psi) &= (-i)^{|\alpha|} D^\alpha F\left(e^{-c|x|^2}\right) = (-i)^{|\alpha|} D^\alpha \left(\left(\frac{1}{2c}\right)^{n/2} e^{-\frac{1}{4c}|s|^2}\right) \\ &= (-i)^{|\alpha|} (-1)^{|\alpha|} \left(\frac{1}{2c}\right)^{|\alpha|+n/2} s^\alpha e^{-\frac{1}{4c}|s|^2} \end{aligned}$$

Similarly, for $\psi(x) = x^\alpha e^{-c|x|^2}$,

$$F^{-1}(\psi) = i^{|\alpha|}(-1)^{|\alpha|}\left(\frac{1}{2c}\right)^{|\alpha|+n/2} s^\alpha e^{-\frac{1}{4c}|s|^2} \tag{13.2}$$

For $\psi(x) = x^\alpha e^{-c|x|^2}$,

$$F^{-1}\circ F(\psi) = F^{-1}\left((-i)^{|\alpha|}(-1)^{|\alpha|}\left(\frac{1}{2c}\right)^{|\alpha|+n/2} s^\alpha e^{-\frac{1}{4c}|s|^2}\right)$$

$$= i^{|\alpha|}\left(\frac{1}{2c}\right)^{|\alpha|+n/2} F^{-1}\left(s^\alpha e^{-\frac{1}{4c}|s|^2}\right)$$

From 13.2 with $c \to 1/(4c)$,

$$= i^{|\alpha|}\left(\frac{1}{2c}\right)^{|\alpha|+n/2} i^{|\alpha|}(-1)^{|\alpha|}\left(\frac{1}{2(1/4c)}\right)^{|\alpha|+n/2} s^\alpha e^{-\frac{1}{4(1/(4c))}|s|^2} = s^\alpha e^{-c|s|^2}$$

Since $\mathcal{G}$ consists of sums of multiples of such ψ, this has shown that $F^{-1}\circ F(\psi) = \psi$. ■

13.2 Fourier Transforms of Just about Anything

I will define Fourier Transforms of the linear functions acting on $\mathcal{G}$ and then show that this includes virtually anything which could possibly be of any interest.

Definition 13.2.1 *Let $\mathcal{G}^*$ denote the vector space of linear functions defined on $\mathcal{G}$ which have values in $\mathbb{C}$. Thus $T \in \mathcal{G}^*$ means $T : \mathcal{G} \to \mathbb{C}$ and T is linear,*

$$T(a\psi + b\phi) = aT(\psi) + bT(\phi) \text{ for all } a, b \in \mathbb{C}, \psi, \phi \in \mathcal{G}$$

Let $\psi \in \mathcal{G}$. Then ψ is an element of $\mathcal{G}^*$ by defining $\psi(\phi) \equiv \int_{\mathbb{R}^n} \psi(x)\phi(x)\,dx$.

Then we have the following important lemma.

Lemma 13.2.2 *The following is obtained for all $\phi, \psi \in \mathcal{G}$.*

$$F\psi(\phi) = \psi(F\phi), \quad F^{-1}\psi(\phi) = \psi\left(F^{-1}\phi\right)$$

Also if $\psi \in \mathcal{G}$ and $\psi = 0$ in $\mathcal{G}^*$ so that $\psi(\phi) = 0$ for all $\phi \in \mathcal{G}$, then $\psi = 0$ as a function.

Proof:

$$F\psi(\phi) \equiv \int_{\mathbb{R}^n} F\psi(t)\phi(t)\,dt = \int_{\mathbb{R}^n}\left(\frac{1}{2\pi}\right)^{n/2}\int_{\mathbb{R}^n} e^{-it\cdot x}\psi(x)dx\phi(t)\,dt$$

$$= \int_{\mathbb{R}^n}\psi(x)\left(\frac{1}{2\pi}\right)^{n/2}\int_{\mathbb{R}^n} e^{-it\cdot x}\phi(t)\,dtdx = \int_{\mathbb{R}^n}\psi(x)F\phi(x)\,dx \equiv \psi(F\phi)$$

The other claim is similar.

Suppose now $\psi(\phi) = 0$ for all $\phi \in \mathcal{G}$. Then $\int_{\mathbb{R}^n}\psi\phi dx = 0$ for all $\phi \in \mathcal{G}$. Therefore, this is true for $\phi = \bar{\psi}$ and so $\psi = 0$. ■

This lemma suggests a way to define the Fourier transform of something in $\mathcal{G}^*$.

Definition 13.2.3 *For $T \in \mathscr{G}^*$, define $FT, F^{-1}T \in \mathscr{G}^*$ by*

$$FT(\phi) \equiv T(F\phi), \; F^{-1}T(\phi) \equiv T\left(F^{-1}\phi\right)$$

Lemma 13.2.4 *F and F^{-1} are both one to one, onto, and are inverses of each other.*

Proof: First note F and F^{-1} are both linear. This follows directly from the definition. Suppose now $FT = 0$. Then $FT(\phi) = T(F\phi) = 0$ for all $\phi \in \mathscr{G}$. But F and F^{-1} map $\mathscr{G}$ onto $\mathscr{G}$ because if $\psi \in \mathscr{G}$, then as shown above, $\psi = F\left(F^{-1}(\psi)\right)$. Therefore, $T = 0$ and so F is one to one. Similarly F^{-1} is one to one. Now

$$F^{-1}(FT)(\phi) \equiv (FT)\left(F^{-1}\phi\right) \equiv T\left(F\left(F^{-1}(\phi)\right)\right) = T\phi.$$

Therefore, $F^{-1} \circ F(T) = T$. Similarly, $F \circ F^{-1}(T) = T$. Thus both F and F^{-1} are one to one and onto and are inverses of each other as suggested by the notation. ■

Probably the most interesting things in $\mathscr{G}^*$ are functions of various kinds. The following lemma will be useful in considering this situation.

Lemma 13.2.5 *If $f \in L^1_{loc}(\mathbb{R}^n)$ and $\int_{\mathbb{R}^n} f\phi dx = 0$ for all $\phi \in C_c(\mathbb{R}^n)$, then $f = 0$ a.e.*

Proof: Let E be bounded and Lebesgue measurable. By regularity, there exists a compact set $K_k \subseteq E$ and an open set $V_k \supseteq E$ such that $m_n(V_k \setminus K_k) < 2^{-k}$. Let h_k equal 1 on K_k, vanish on V_k^C, and take values between 0 and 1. Let $E_k \equiv \left[\left|h_k - \mathscr{X}_{E_k}\right| > \left(\frac{2}{3}\right)^k\right]$. Then

$$m_n(E_k) \leq \left(\frac{3}{2}\right)^k \int_{E_k} \left|h_k - \mathscr{X}_{E_k}\right| dm_n < \left(\frac{3}{2}\right)^k \int_{V_k \setminus K_k} 2dm_n = 2\left(\frac{3}{2}\right)^k \frac{1}{2^k} = 2\left(\frac{3}{4}\right)^k$$

and so $\sum_k m_n(E_k) < \infty$ so there is a set of measure zero such that off this set, $h_k \to \mathscr{X}_{E_k}$.

Hence, by the dominated convergence theorem, $\int f\mathscr{X}_E dm_n = \lim_{k\to\infty} \int f h_k dm_n = 0$. It follows that for E an arbitrary Lebesgue measurable set, $\int f\mathscr{X}_{B(\mathbf{0},R)}\mathscr{X}_E dm_n = 0$. Let

$$\operatorname{sgn} f = \begin{cases} \frac{\overline{f}}{|f|} \text{ if } |f| \neq 0 \\ 0 \text{ if } |f| = 0 \end{cases}$$

By Theorem 9.1.6 applied to positive and negative parts, there exists $\{s_k\}$, a sequence of simple functions converging pointwise to $\operatorname{sgn} f$ such that $|s_k| \leq 1$. Then by the dominated convergence theorem again, $\int |f| \mathscr{X}_{B(\mathbf{0},R)} dm_n = \lim_{k\to\infty} \int f\mathscr{X}_{B(\mathbf{0},R)} s_k dm_n = 0$. Since R is arbitrary, $|f| = 0$ a.e. ■

Corollary 13.2.6 *Let $f \in L^1(\mathbb{R}^n)$ and suppose $\int_{\mathbb{R}^n} f(\boldsymbol{x})\phi(\boldsymbol{x})dx = 0$ for all $\phi \in \mathscr{G}$. Then $f = 0$ a.e.*

Proof: Let $\psi \in C_c(\mathbb{R}^n)$. Then by the Stone Weierstrass approximation theorem, there exists a sequence of functions, $\{\phi_k\} \subseteq \mathscr{G}$ such that $\phi_k \to \psi$ uniformly. Then by the dominated convergence theorem, $\int f\psi dx = \lim_{k\to\infty} \int f\phi_k dx = 0$. By Lemma 13.2.5 $f = 0$. ■

The next theorem is the main result of this sort.

Theorem 13.2.7 *Let $f \in L^p(\mathbb{R}^n)$, $p \geq 1$, or suppose f is measurable and has polynomial growth, $|f(\boldsymbol{x})| \leq K\left(1 + |\boldsymbol{x}|^2\right)^m$ for some $m \in \mathbb{N}$. Then if $\int f\psi dx = 0$, for all $\psi \in \mathscr{G}$, then it follows $f = 0$.*

Proof: First note that if $f \in L^p(\mathbb{R}^n)$ or has polynomial growth, then it makes sense to write the integral $\int f\psi dx$ described above. This is obvious in the case of polynomial growth. In the case where $f \in L^p(\mathbb{R}^n)$ it also makes sense because

$$\int |f||\psi| dx \leq \left(\int |f|^p dx\right)^{1/p} \left(\int |\psi|^{p'} dx\right)^{1/p'} < \infty$$

due to the fact mentioned above that all these functions in $\mathscr{G}$ are in $L^p(\mathbb{R}^n)$ for every $p \geq 1$. Suppose now that $f \in L^p, p \geq 1$. The case where $f \in L^1(\mathbb{R}^n)$ was dealt with in Corollary 13.2.6. Suppose $f \in L^p(\mathbb{R}^n)$ for $p > 1$. Then

$$|f|^{p-2}\overline{f} \in L^{p'}(\mathbb{R}^n), \left(p' = q, \frac{1}{p} + \frac{1}{q} = 1\right)$$

and by density of $\mathscr{G}$ in $L^{p'}(\mathbb{R}^n)$ (Theorem 13.1.4), there exists a sequence $\{g_k\} \subseteq \mathscr{G}$ such that $\left\|g_k - |f|^{p-2}\overline{f}\right\|_{p'} \to 0$. Then

$$\begin{aligned}\int_{\mathbb{R}^n} |f|^p dx &= \int_{\mathbb{R}^n} f\left(|f|^{p-2}\overline{f} - g_k\right) dx + \int_{\mathbb{R}^n} f g_k dx \\ &= \int_{\mathbb{R}^n} f\left(|f|^{p-2}\overline{f} - g_k\right) dx \leq \|f\|_{L^p} \left\|g_k - |f|^{p-2}\overline{f}\right\|_{p'}\end{aligned}$$

which converges to 0. Hence $f = 0$.

It remains to consider the case where f has polynomial growth. Thus $\boldsymbol{x} \to f(\boldsymbol{x}) e^{-|\boldsymbol{x}|^2} \in L^1(\mathbb{R}^n)$. Therefore, for all $\psi \in \mathscr{G}$, $0 = \int f(\boldsymbol{x}) e^{-|\boldsymbol{x}|^2} \psi(\boldsymbol{x}) dx$ because $e^{-|\boldsymbol{x}|^2} \psi(\boldsymbol{x}) \in \mathscr{G}$. Therefore, by the first part, $f(\boldsymbol{x}) e^{-|\boldsymbol{x}|^2} = 0$ a.e. ■

Note that "polynomial growth" could be replaced with a condition of the form

$$|f(\boldsymbol{x})| \leq K\left(1 + |\boldsymbol{x}|^2\right)^m e^{k|x|^\alpha}, \ \alpha < 2$$

and the same proof would yield that these functions are in $\mathscr{G}^*$. The main thing to observe is that almost all functions of interest are in $\mathscr{G}^*$.

Theorem 13.2.8

Let f be a measurable function with polynomial growth,

$$|f(\boldsymbol{x})| \leq C\left(1 + |\boldsymbol{x}|^2\right)^N \text{ for some } N,$$

or let $f \in L^p(\mathbb{R}^n)$ for some $p \in [1, \infty]$. Then $f \in \mathscr{G}^*$ if $f(\phi) \equiv \int f\phi dx$.

Proof: Let f have polynomial growth first. Then the above integral is clearly well defined and so in this case, $f \in \mathscr{G}^*$.

Next suppose $f \in L^p(\mathbb{R}^n)$ with $\infty > p \geq 1$. Then it is clear again that the above integral is well defined because of the fact that ϕ is a sum of polynomials times exponentials of the form $e^{-c|\boldsymbol{x}|^2}$ and these are in $L^{p'}(\mathbb{R}^n)$. Also $\phi \to f(\phi)$ is clearly linear in both cases. ■

This has shown that for nearly any reasonable function, you can define its Fourier transform as described above. You could also define the Fourier transform of a finite Borel measure μ because for such a measure $\psi \to \int_{\mathbb{R}^n} \psi d\mu$ is a linear functional on $\mathscr{G}$. This includes the very important case of probability distribution measures. The theoretical basis for this assertion will be given a little later.

13.2.1 Fourier Transforms of Functions in $L^1(\mathbb{R}^n)$

First suppose $f \in L^1(\mathbb{R}^n)$.

Theorem 13.2.9 *Let $f \in L^1(\mathbb{R}^n)$. Then $Ff(\phi) = \int_{\mathbb{R}^n} g\phi dt$ where*

$$g(\boldsymbol{t}) = \left(\frac{1}{2\pi}\right)^{n/2} \int_{\mathbb{R}^n} e^{-i\boldsymbol{t}\cdot\boldsymbol{x}} f(\boldsymbol{x})\, dx$$

and $F^{-1}f(\phi) = \int_{\mathbb{R}^n} g\phi dt$ where $g(\boldsymbol{t}) = \left(\frac{1}{2\pi}\right)^{n/2} \int_{\mathbb{R}^n} e^{i\boldsymbol{t}\cdot\boldsymbol{x}} f(\boldsymbol{x})\, dx$. In short,

$$Ff(\boldsymbol{t}) \equiv (2\pi)^{-n/2} \int_{\mathbb{R}^n} e^{-i\boldsymbol{t}\cdot\boldsymbol{x}} f(\boldsymbol{x}) dx,\ F^{-1}f(\boldsymbol{t}) \equiv (2\pi)^{-n/2} \int_{\mathbb{R}^n} e^{i\boldsymbol{t}\cdot\boldsymbol{x}} f(\boldsymbol{x}) dx.$$

Proof: From the definition and Fubini's theorem,

$$\begin{aligned} Ff(\phi) &\equiv \int_{\mathbb{R}^n} f(\boldsymbol{t}) F\phi(\boldsymbol{t})\, dt = \int_{\mathbb{R}^n} f(\boldsymbol{t}) \left(\frac{1}{2\pi}\right)^{n/2} \int_{\mathbb{R}^n} e^{-i\boldsymbol{t}\cdot\boldsymbol{x}} \phi(\boldsymbol{x})\, dx dt \\ &= \int_{\mathbb{R}^n} \left(\left(\frac{1}{2\pi}\right)^{n/2} \int_{\mathbb{R}^n} f(\boldsymbol{t}) e^{-i\boldsymbol{t}\cdot\boldsymbol{x}} dt \right) \phi(\boldsymbol{x})\, dx. \end{aligned}$$

Since $\phi \in \mathscr{G}$ is arbitrary, it follows from Theorem 13.2.7 that $Ff(\boldsymbol{x})$ is given by the claimed formula. The case of F^{-1} is identical. ■

Here are interesting properties of these Fourier transforms of functions in L^1.

Theorem 13.2.10 *If $f \in L^1(\mathbb{R}^n)$ and $\|f_k - f\|_1 \to 0$, then Ff_k and $F^{-1}f_k$ converge uniformly to Ff and $F^{-1}f$ respectively. If $f \in L^1(\mathbb{R}^n)$, then $F^{-1}f$ and Ff are both continuous and bounded. Also,*

$$\lim_{|\boldsymbol{x}|\to\infty} F^{-1}f(\boldsymbol{x}) = \lim_{|\boldsymbol{x}|\to\infty} Ff(\boldsymbol{x}) = 0. \tag{13.3}$$

Furthermore, for $f \in L^1(\mathbb{R}^n)$ both Ff and $F^{-1}f$ are uniformly continuous.

Proof: The first claim follows from the following inequality.

$$\begin{aligned} |Ff_k(\boldsymbol{t}) - Ff(\boldsymbol{t})| &\leq (2\pi)^{-n/2} \int_{\mathbb{R}^n} \left| e^{-i\boldsymbol{t}\cdot\boldsymbol{x}} f_k(\boldsymbol{x}) - e^{-i\boldsymbol{t}\cdot\boldsymbol{x}} f(\boldsymbol{x}) \right| dx \\ &= (2\pi)^{-n/2} \int_{\mathbb{R}^n} |f_k(\boldsymbol{x}) - f(\boldsymbol{x})|\, dx = (2\pi)^{-n/2} \|f - f_k\|_1. \end{aligned}$$

which a similar argument holding for F^{-1}.

Now consider the second claim of the theorem.

$$\left|Ff(\boldsymbol{t}) - Ff(\boldsymbol{t}')\right| \leq (2\pi)^{-n/2} \int_{\mathbb{R}^n} \left| e^{-i\boldsymbol{t}\cdot\boldsymbol{x}} - e^{-i\boldsymbol{t}'\cdot\boldsymbol{x}} \right| |f(\boldsymbol{x})|\, dx$$

The integrand is bounded by $2|f(\boldsymbol{x})|$, a function in $L^1(\mathbb{R}^n)$ and converges to 0 as $\boldsymbol{t}' \to \boldsymbol{t}$ and so the dominated convergence theorem implies Ff is continuous. To see $Ff(\boldsymbol{t})$ is uniformly bounded,

$$|Ff(\boldsymbol{t})| \leq (2\pi)^{-n/2} \int_{\mathbb{R}^n} |f(\boldsymbol{x})|\, dx < \infty.$$

A similar argument gives the same conclusions for F^{-1}.

It remains to verify 13.3 and the claim that Ff and $F^{-1}f$ are uniformly continuous.

$$|Ff(\boldsymbol{t})| \leq \left|(2\pi)^{-n/2}\int_{\mathbb{R}^n} e^{-i\boldsymbol{t}\cdot\boldsymbol{x}} f(\boldsymbol{x})dx\right|$$

Now let $\varepsilon > 0$ be given and let $g \in C_c^\infty(\mathbb{R}^n)$ such that $(2\pi)^{-n/2}\|g-f\|_1 < \varepsilon/2$. Then

$$\begin{aligned} |Ff(\boldsymbol{t})| &\leq (2\pi)^{-n/2}\int_{\mathbb{R}^n} |f(\boldsymbol{x})-g(\boldsymbol{x})|\,dx + \left|(2\pi)^{-n/2}\int_{\mathbb{R}^n} e^{-i\boldsymbol{t}\cdot\boldsymbol{x}} g(\boldsymbol{x})dx\right| \\ &\leq \varepsilon/2 + \left|(2\pi)^{-n/2}\int_{\mathbb{R}^n} e^{-i\boldsymbol{t}\cdot\boldsymbol{x}} g(\boldsymbol{x})dx\right|. \end{aligned}$$

Now integrating by parts, it follows that for $\|\boldsymbol{t}\|_\infty \equiv \max\{|t_j| : j = 1,\cdots,n\} > 0$

$$|Ff(\boldsymbol{t})| \leq \varepsilon/2 + (2\pi)^{-n/2}\left|\frac{1}{\|\boldsymbol{t}\|_\infty}\int_{\mathbb{R}^n}\sum_{j=1}^n \left|\frac{\partial g(\boldsymbol{x})}{\partial x_j}\right| dx\right| \tag{13.4}$$

and this last expression converges to zero as $\|\boldsymbol{t}\|_\infty \to \infty$. The reason for this is that if $t_j \neq 0$, integration by parts with respect to x_j gives

$$(2\pi)^{-n/2}\int_{\mathbb{R}^n} e^{-i\boldsymbol{t}\cdot\boldsymbol{x}} g(\boldsymbol{x})dx = (2\pi)^{-n/2}\frac{1}{-it_j}\int_{\mathbb{R}^n} e^{-i\boldsymbol{t}\cdot\boldsymbol{x}}\frac{\partial g(\boldsymbol{x})}{\partial x_j}dx.$$

Therefore, choose the j for which $\|\boldsymbol{t}\|_\infty = |t_j|$ and the result of 13.4 holds. Therefore, from 13.4, if $\|\boldsymbol{t}\|_\infty$ is large enough, $|Ff(\boldsymbol{t})| < \varepsilon$. Similarly, $\lim_{\|\boldsymbol{t}\|\to\infty} F^{-1}(\boldsymbol{t}) = 0$.

Consider the claim about uniform continuity. Let $\varepsilon > 0$ be given. Then there exists R such that if $\|\boldsymbol{t}\|_\infty > R$, then $|Ff(\boldsymbol{t})| < \frac{\varepsilon}{2}$. Since Ff is continuous, it is uniformly continuous on the compact set $[-R-1, R+1]^n$. Therefore, there exists δ_1 such that if $\|\boldsymbol{t}-\boldsymbol{t}'\|_\infty < \delta_1$ for $\boldsymbol{t}', \boldsymbol{t} \in [-R-1, R+1]^n$, then

$$|Ff(\boldsymbol{t}) - Ff(\boldsymbol{t}')| < \varepsilon/2. \tag{13.5}$$

Now let $0 < \delta < \min(\delta_1, 1)$ and suppose $\|\boldsymbol{t}-\boldsymbol{t}'\|_\infty < \delta$. If both $\boldsymbol{t}, \boldsymbol{t}'$ are contained in $[-R,R]^n$, then 13.5 holds. If $\boldsymbol{t} \in [-R,R]^n$ and $\boldsymbol{t}' \notin [-R,R]^n$, then both are contained in $[-R-1, R+1]^n$ and so this verifies 13.5 in this case. The other case is that neither point is in $[-R,R]^n$ and in this case,

$$|Ff(\boldsymbol{t}) - Ff(\boldsymbol{t}')| \leq |Ff(\boldsymbol{t})| + |Ff(\boldsymbol{t}')| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \ \blacksquare$$

There is a very interesting relation between the Fourier transform and convolutions.

Theorem 13.2.11 *Let $f, g \in L^1(\mathbb{R}^n)$. Then $f * g \in L^1$, $F(f*g) = (2\pi)^{n/2} FfFg$.*

Proof: Consider

$$\int_{\mathbb{R}^n}\int_{\mathbb{R}^n} |f(\boldsymbol{x}-\boldsymbol{y})\,g(\boldsymbol{y})|\,dydx.$$

The function, $(\boldsymbol{x},\boldsymbol{y}) \to |f(\boldsymbol{x}-\boldsymbol{y})\,g(\boldsymbol{y})|$ is Lebesgue measurable and so by Fubini's theorem,

$$\int_{\mathbb{R}^n}\int_{\mathbb{R}^n} |f(\boldsymbol{x}-\boldsymbol{y})\,g(\boldsymbol{y})|\,dydx = \int_{\mathbb{R}^n}\int_{\mathbb{R}^n} |f(\boldsymbol{x}-\boldsymbol{y})\,g(\boldsymbol{y})|\,dxdy = \|f\|_1\|g\|_1 < \infty.$$

It follows that for a.e. $\boldsymbol{x}$, $\int_{\mathbb{R}^n} |f(\boldsymbol{x}-\boldsymbol{y})\, g(\boldsymbol{y})|\, dy < \infty$ and for each of these values of $\boldsymbol{x}$, it follows that $\int_{\mathbb{R}^n} f(\boldsymbol{x}-\boldsymbol{y})\, g(\boldsymbol{y})\, dy$ exists and equals a function of $\boldsymbol{x}$ which is in $L^1(\mathbb{R}^n)$, $f * g(\boldsymbol{x})$. Now

$$\begin{aligned} & F(f*g)(\boldsymbol{t}) \equiv (2\pi)^{-n/2} \int_{\mathbb{R}^n} e^{-i\boldsymbol{t}\cdot\boldsymbol{x}} f*g(\boldsymbol{x})\, dx \\ = \; & (2\pi)^{-n/2} \int_{\mathbb{R}^n} e^{-i\boldsymbol{t}\cdot\boldsymbol{x}} \int_{\mathbb{R}^n} f(\boldsymbol{x}-\boldsymbol{y})\, g(\boldsymbol{y})\, dy dx \\ = \; & (2\pi)^{-n/2} \int_{\mathbb{R}^n} e^{-i\boldsymbol{t}\cdot\boldsymbol{y}} g(\boldsymbol{y}) \int_{\mathbb{R}^n} e^{-i\boldsymbol{t}\cdot(\boldsymbol{x}-\boldsymbol{y})} f(\boldsymbol{x}-\boldsymbol{y})\, dx dy \\ = \; & (2\pi)^{n/2} Ff(\boldsymbol{t})\, Fg(\boldsymbol{t}). \blacksquare \end{aligned}$$

There are many other considerations involving Fourier transforms of functions in L^1. Some others are in the exercises.

13.2.2 Fourier Transforms of Functions in $L^2(\mathbb{R}^n)$

Consider Ff and $F^{-1}f$ for $f \in L^2(\mathbb{R}^n)$. First note that the formula given for Ff and $F^{-1}f$ when $f \in L^1(\mathbb{R}^n)$ will not work for $f \in L^2(\mathbb{R}^n)$ unless f is also in $L^1(\mathbb{R}^n)$. Recall that $\overline{a+ib} = a-ib$.

Theorem 13.2.12 *For $\phi \in \mathcal{G}$, $\|F\phi\|_2 = \|F^{-1}\phi\|_2 = \|\phi\|_2$.*

Proof: First note that for $\psi \in \mathcal{G}$,

$$F(\overline{\psi}) = \overline{F^{-1}(\psi)}\,,\; F^{-1}(\overline{\psi}) = \overline{F(\psi)}. \tag{13.6}$$

This follows from the definition. For example,

$$F\overline{\psi}(\boldsymbol{t}) = (2\pi)^{-n/2} \int_{\mathbb{R}^n} e^{-i\boldsymbol{t}\cdot\boldsymbol{x}} \overline{\psi}(\boldsymbol{x})\, dx = \overline{(2\pi)^{-n/2} \int_{\mathbb{R}^n} e^{i\boldsymbol{t}\cdot\boldsymbol{x}} \psi(\boldsymbol{x})\, dx}$$

Let $\phi, \psi \in \mathcal{G}$. It was shown above that $\int_{\mathbb{R}^n} (F\phi)\psi(\boldsymbol{t})dt = \int_{\mathbb{R}^n} \phi(F\psi)dx$. Similarly,

$$\int_{\mathbb{R}^n} \phi(F^{-1}\psi)dx = \int_{\mathbb{R}^n} (F^{-1}\phi)\psi dt. \tag{13.7}$$

Now, 13.6 - 13.7 imply

$$\begin{aligned} \int_{\mathbb{R}^n} |\phi|^2 dx &= \int_{\mathbb{R}^n} \phi\overline{\phi} dx = \int_{\mathbb{R}^n} \phi(\overline{F^{-1}(F\phi)})dx = \int_{\mathbb{R}^n} \phi F(\overline{F\phi})dx \\ &= \int_{\mathbb{R}^n} F\phi(\overline{F\phi})dx = \int_{\mathbb{R}^n} |F\phi|^2 dx. \end{aligned}$$

Similarly $\|\phi\|_2 = \|F^{-1}\phi\|_2$. $\blacksquare$

Lemma 13.2.13 *Let $f \in L^2(\mathbb{R}^n)$ and let $\phi_k \to f$ in $L^2(\mathbb{R}^n)$ where $\phi_k \in \mathcal{G}$. (Such a sequence exists because of density of $\mathcal{G}$ in $L^2(\mathbb{R}^n)$.) Then Ff and $F^{-1}f$ are both in $L^2(\mathbb{R}^n)$ and the following limits take place in L^2.*

$$\lim_{k\to\infty} F(\phi_k) = F(f),\; \lim_{k\to\infty} F^{-1}(\phi_k) = F^{-1}(f).$$

Proof: Let $\psi \in \mathscr{G}$ be given. Then from Theorem 13.2.8,

$$\begin{aligned} Ff(\psi) &\equiv f(F\psi) \equiv \int_{\mathbb{R}^n} f(\boldsymbol{x}) F\psi(\boldsymbol{x})\, dx \\ &= \lim_{k\to\infty} \int_{\mathbb{R}^n} \phi_k(\boldsymbol{x}) F\psi(\boldsymbol{x})\, dx = \lim_{k\to\infty} \int_{\mathbb{R}^n} F\phi_k(\boldsymbol{x})\, \psi(\boldsymbol{x})\, dx. \end{aligned}$$

Also by Theorem 13.2.12 $\{F\phi_k\}_{k=1}^{\infty}$ is Cauchy in $L^2(\mathbb{R}^n)$ since

$$\|F\phi_k - F\phi_l\|_{L^2} = \|\phi_k - \phi_l\|_{L^2},$$

and so $\lim_{k\to\infty} F\phi_k = h$ for some $h \in L^2(\mathbb{R}^n)$. Therefore, from the above, $Ff(\psi) = \int_{\mathbb{R}^n} h(\boldsymbol{x})\, \psi(\boldsymbol{x})$ which shows that $F(f) \in L^2(\mathbb{R}^n)$ and $h = F(f)$. The case of F^{-1} is entirely similar. ■

Since Ff and $F^{-1}f$ are in $L^2(\mathbb{R}^n)$, this also proves the following theorem.

Theorem 13.2.14 *If $f \in L^2(\mathbb{R}^n)$, Ff and $F^{-1}f$ are the unique elements of $L^2(\mathbb{R}^n)$ such that for all $\phi \in \mathscr{G}$,*

$$\int_{\mathbb{R}^n} Ff(\boldsymbol{x})\phi(\boldsymbol{x})dx = \int_{\mathbb{R}^n} f(\boldsymbol{x})F\phi(\boldsymbol{x})dx, \tag{13.8}$$

$$\int_{\mathbb{R}^n} F^{-1}f(\boldsymbol{x})\phi(\boldsymbol{x})dx = \int_{\mathbb{R}^n} f(\boldsymbol{x})F^{-1}\phi(\boldsymbol{x})dx. \tag{13.9}$$

Theorem 13.2.15 *(Plancherel)*

$$\|f\|_2 = \|Ff\|_2 = \|F^{-1}f\|_2. \tag{13.10}$$

Proof: Use the density of $\mathscr{G}$ in $L^2(\mathbb{R}^n)$ to obtain a sequence, $\{\phi_k\}$ converging to f in $L^2(\mathbb{R}^n)$. Then by Lemma 13.2.13

$$\|Ff\|_2 = \lim_{k\to\infty} \|F\phi_k\|_2 = \lim_{k\to\infty} \|\phi_k\|_2 = \|f\|_2.$$

Similarly, $\|f\|_2 = \|F^{-1}f\|_2$. ■

The following corollary is a generalization of this. To prove this corollary, use the following simple lemma which comes as a consequence of the Cauchy Schwarz inequality.

Lemma 13.2.16 *Suppose $f_k \to f$ in $L^2(\mathbb{R}^n)$ and $g_k \to g$ in $L^2(\mathbb{R}^n)$. Then*

$$\lim_{k\to\infty} \int_{\mathbb{R}^n} f_k g_k dx = \int_{\mathbb{R}^n} fg dx$$

Proof:

$$\left|\int_{\mathbb{R}^n} f_k g_k dx - \int_{\mathbb{R}^n} fg dx\right| \le \left|\int_{\mathbb{R}^n} f_k g_k dx - \int_{\mathbb{R}^n} f_k g dx\right| + \left|\int_{\mathbb{R}^n} f_k g dx - \int_{\mathbb{R}^n} fg dx\right|$$

$$\le \|f_k\|_2 \|g - g_k\|_2 + \|g\|_2 \|f_k - f\|_2.$$

Now $\|f_k\|_2$ is a Cauchy sequence and so it is bounded independent of k. Therefore, the above expression is smaller than ε whenever k is large enough. ■

Corollary 13.2.17 *For $f,g \in L^2(\mathbb{R}^n)$,*

$$\int_{\mathbb{R}^n} f\overline{g}dx = \int_{\mathbb{R}^n} Ff\,\overline{Fg}dx = \int_{\mathbb{R}^n} F^{-1}f\,\overline{F^{-1}g}dx.$$

Proof: First note the above formula is obvious if $f,g \in \mathcal{G}$. To see this, note

$$\int_{\mathbb{R}^n} Ff\,\overline{Fg}dx = \int_{\mathbb{R}^n} Ff(\boldsymbol{x})\,\overline{\frac{1}{(2\pi)^{n/2}}\int_{\mathbb{R}^n} e^{-i\boldsymbol{x}\cdot\boldsymbol{t}}g(\boldsymbol{t})\,dt}dx$$

$$= \int_{\mathbb{R}^n} \frac{1}{(2\pi)^{n/2}}\int_{\mathbb{R}^n} e^{i\boldsymbol{x}\cdot\boldsymbol{t}}Ff(\boldsymbol{x})\,dx\overline{g(\boldsymbol{t})}dt = \int_{\mathbb{R}^n} \left(F^{-1}\circ F\right)f(\boldsymbol{t})\,\overline{g(\boldsymbol{t})}dt = \int_{\mathbb{R}^n} f(\boldsymbol{t})\,\overline{g(\boldsymbol{t})}dt.$$

The formula with F^{-1} is exactly similar.

Now to verify the corollary, let $\phi_k \to f$ in $L^2(\mathbb{R}^n)$ and let $\psi_k \to g$ in $L^2(\mathbb{R}^n)$. Then by Lemma 13.2.13

$$\int_{\mathbb{R}^n} Ff\,\overline{Fg}dx = \lim_{k\to\infty}\int_{\mathbb{R}^n} F\phi_k\,\overline{F\psi_k}dx = \lim_{k\to\infty}\int_{\mathbb{R}^n} \phi_k\overline{\psi_k}dx = \int_{\mathbb{R}^n} f\overline{g}dx$$

A similar argument holds for F^{-1}. ■

How does one compute Ff and $F^{-1}f$?

Theorem 13.2.18 *For $f \in L^2(\mathbb{R}^n)$, let $f_r = f\mathcal{X}_{E_r}$ where E_r is a bounded measurable set with $E_r \uparrow \mathbb{R}^n$. Then the following limits hold in $L^2(\mathbb{R}^n)$.*

$$Ff = \lim_{r\to\infty} Ff_r\ ,\ F^{-1}f = \lim_{r\to\infty} F^{-1}f_r.$$

Proof: $\|f - f_r\|_2 \to 0$ and so $\|Ff - Ff_r\|_2 \to 0$ and $\|F^{-1}f - F^{-1}f_r\|_2 \to 0$ which both follow from Plancherel's Theorem. ■

What are Ff_r and $F^{-1}f_r$? Let $\phi \in \mathcal{G}$

$$\begin{aligned}\int_{\mathbb{R}^n} Ff_r\phi dx &= \int_{\mathbb{R}^n} f_rF\phi dx = (2\pi)^{-\frac{n}{2}}\int_{\mathbb{R}^n}\int_{\mathbb{R}^n} f_r(\boldsymbol{x})e^{-i\boldsymbol{x}\cdot\boldsymbol{y}}\phi(\boldsymbol{y})dydx \\ &= \int_{\mathbb{R}^n}[(2\pi)^{-\frac{n}{2}}\int_{\mathbb{R}^n} f_r(\boldsymbol{x})e^{-i\boldsymbol{x}\cdot\boldsymbol{y}}dx]\phi(\boldsymbol{y})dy.\end{aligned}$$

Since this holds for all $\phi \in \mathcal{G}$, a dense subset of $L^2(\mathbb{R}^n)$, it follows that

$$Ff_r(\boldsymbol{y}) = (2\pi)^{-\frac{n}{2}}\int_{\mathbb{R}^n} f_r(\boldsymbol{x})e^{-i\boldsymbol{x}\cdot\boldsymbol{y}}dx.$$

Similarly

$$F^{-1}f_r(\boldsymbol{y}) = (2\pi)^{-\frac{n}{2}}\int_{\mathbb{R}^n} f_r(\boldsymbol{x})e^{i\boldsymbol{x}\cdot\boldsymbol{y}}dx.$$

This shows that to take the Fourier transform of a function in $L^2(\mathbb{R}^n)$, it suffices to take the limit as $r\to\infty$ in $L^2(\mathbb{R}^n)$ of $(2\pi)^{-\frac{n}{2}}\int_{\mathbb{R}^n} f_r(\boldsymbol{x})e^{-i\boldsymbol{x}\cdot\boldsymbol{y}}dx$. A similar procedure works for the inverse Fourier transform.

Note this reduces to the earlier definition in case $f \in L^1(\mathbb{R}^n)$. Now consider the convolution of a function in L^2 with one in L^1.

Theorem 13.2.19 *Let $h \in L^2(\mathbb{R}^n)$ and let $f \in L^1(\mathbb{R}^n)$. Then $h*f \in L^2(\mathbb{R}^n)$,*

$$F^{-1}(h*f) = (2\pi)^{n/2} F^{-1}hF^{-1}f,\ F(h*f) = (2\pi)^{n/2} FhFf,$$

and

$$||h*f||_2 \leq \|h\|_2 \|f\|_1. \tag{13.11}$$

Proof: An application of Minkowski's inequality yields

$$\left(\int_{\mathbb{R}^n}\left(\int_{\mathbb{R}^n}|h(\boldsymbol{x}-\boldsymbol{y})||f(\boldsymbol{y})|\,dy\right)^2 dx\right)^{1/2} \leq \|f\|_1 \|h\|_2. \tag{13.12}$$

Hence $\int |h(\boldsymbol{x}-\boldsymbol{y})||f(\boldsymbol{y})|\,dy < \infty$ a.e. $\boldsymbol{x}$ and $\boldsymbol{x} \to \int h(\boldsymbol{x}-\boldsymbol{y}) f(\boldsymbol{y})\,dy$ is in $L^2(\mathbb{R}^n)$. Let $E_r \uparrow \mathbb{R}^n$, $m(E_r) < \infty$. Thus, $h_r \equiv \mathcal{X}_{E_r} h \in L^2(\mathbb{R}^n) \cap L^1(\mathbb{R}^n)$, and letting $\phi \in \mathcal{G}$,

$$\int F(h_r * f)(\phi)\,dx =$$

$$\begin{aligned}
&\equiv \int (h_r * f)(F\phi)\,dx = (2\pi)^{-n/2} \int\int\int h_r(\boldsymbol{x}-\boldsymbol{y}) f(\boldsymbol{y}) e^{-i\boldsymbol{x}\cdot\boldsymbol{t}} \phi(\boldsymbol{t})\,dtdydx \\
&= (2\pi)^{-n/2} \int\int \left(\int h_r(\boldsymbol{x}-\boldsymbol{y}) e^{-i(\boldsymbol{x}-\boldsymbol{y})\cdot\boldsymbol{t}} dx\right) f(\boldsymbol{y}) e^{-i\boldsymbol{y}\cdot\boldsymbol{t}} dy \phi(\boldsymbol{t})\,dt \\
&= \int (2\pi)^{n/2} Fh_r(\boldsymbol{t}) Ff(\boldsymbol{t}) \phi(\boldsymbol{t})\,dt.
\end{aligned}$$

Since ϕ is arbitrary and $\mathcal{G}$ is dense in $L^2(\mathbb{R}^n)$, $F(h_r*f) = (2\pi)^{n/2} Fh_r Ff$. Now by Minkowski's Inequality, $h_r * f \to h*f$ in $L^2(\mathbb{R}^n)$ and also it is clear that $h_r \to h$ in $L^2(\mathbb{R}^n)$; so, by Plancherel's theorem, you may take the limit in the above and conclude the following equation: $F(h*f) = (2\pi)^{n/2} FhFf$. The assertion for F^{-1} is similar and 13.11 follows from 13.12. ■

13.2.3 The Schwartz Class

The problem with $\mathcal{G}$ is that it does not contain $C_c^\infty(\mathbb{R}^n)$. I have used it in presenting the Fourier transform because the functions in $\mathcal{G}$ have a very specific form which made some technical details work out easier than in any other approach I have seen. The Schwartz class is a larger class of functions which does contain $C_c^\infty(\mathbb{R}^n)$ and also has the same nice properties as $\mathcal{G}$. The functions in the Schwartz class are infinitely differentiable and they vanish very rapidly as $|\boldsymbol{x}| \to \infty$ along with all their partial derivatives. This is the description of these functions, not a specific form involving polynomials times $e^{-\alpha|\boldsymbol{x}|^2}$. To describe this precisely requires some notation.

Definition 13.2.20 *$f \in \mathcal{S}$, the Schwartz class, if $f \in C^\infty(\mathbb{R}^n)$ and for all positive integers N, $\rho_N(f) < \infty$ where*

$$\rho_N(f) = \sup\{(1+|\boldsymbol{x}|^2)^N |D^\alpha f(\boldsymbol{x})| : \boldsymbol{x} \in \mathbb{R}^n,\ |\alpha| \leq N\}.$$

Thus $f \in \mathcal{S}$ if and only if $f \in C^\infty(\mathbb{R}^n)$ and

$$\sup\{|\boldsymbol{x}^\beta D^\alpha f(\boldsymbol{x})| : \boldsymbol{x} \in \mathbb{R}^n\} < \infty \tag{13.13}$$

for all multi indices α and β.

Also note that if $f \in \mathcal{S}$, then $p(f) \in \mathcal{S}$ for any polynomial, p with $p(0) = 0$ and that

$$\mathcal{S} \subseteq L^p(\mathbb{R}^n) \cap L^\infty(\mathbb{R}^n)$$

for any $p \geq 1$. To see this assertion about the $p(f)$, it suffices to consider the case of the product of two elements of the Schwartz class. If $f, g \in \mathcal{S}$, then $D^\alpha(fg)$ is a finite sum of derivatives of f times derivatives of g. Therefore, $\rho_N(fg) < \infty$ for all N. You may wonder about examples of things in $\mathcal{S}$. Clearly any function in $C_c^\infty(\mathbb{R}^n)$ is in $\mathcal{S}$. However there are other functions in $\mathcal{S}$. For example $e^{-|x|^2}$ is in $\mathcal{S}$ as you can verify for yourself and so is any function from $\mathcal{G}$. Note also that the density of $C_c(\mathbb{R}^n)$ in $L^p(\mathbb{R}^n)$ shows that $\mathcal{S}$ is dense in $L^p(\mathbb{R}^n)$ for every p.

Recall the Fourier transform of a function in $L^1(\mathbb{R}^n)$ is given by

$$Ff(t) \equiv (2\pi)^{-n/2} \int_{\mathbb{R}^n} e^{-it\cdot x} f(x) dx.$$

Therefore, this gives the Fourier transform for $f \in \mathcal{S}$. The nice property which $\mathcal{S}$ has in common with $\mathcal{G}$ is that the Fourier transform and its inverse map $\mathcal{S}$ one to one onto $\mathcal{S}$. This means I could have presented the whole of the above theory in terms of $\mathcal{S}$ rather than in terms of $\mathcal{G}$. However, it is more technical.

Theorem 13.2.21 *If $f \in \mathcal{S}$, then Ff and $F^{-1}f$ are also in $\mathcal{S}$.*

Proof: To begin with, let $\alpha = e_j = (0,0,\cdots,1,0,\cdots,0)$, the 1 in the j^{th} slot.

$$\frac{F^{-1}f(t+he_j) - F^{-1}f(t)}{h} = (2\pi)^{-n/2} \int_{\mathbb{R}^n} e^{it\cdot x} f(x) \left(\frac{e^{ihx_j} - 1}{h} \right) dx. \tag{13.14}$$

Consider the integrand in 13.14.

$$\begin{aligned} \left| e^{it\cdot x} f(x) \left(\frac{e^{ihx_j} - 1}{h} \right) \right| &= |f(x)| \left| \left(\frac{e^{i(h/2)x_j} - e^{-i(h/2)x_j}}{h} \right) \right| \\ &= |f(x)| \left| \frac{i \sin((h/2)x_j)}{(h/2)} \right| \leq |f(x)| \, |x_j| \end{aligned}$$

and this is a function in $L^1(\mathbb{R}^n)$ because $f \in \mathcal{S}$. Therefore by the Dominated Convergence Theorem,

$$\frac{\partial F^{-1}f(t)}{\partial t_j} = (2\pi)^{-n/2} \int_{\mathbb{R}^n} e^{it\cdot x} i x_j f(x) \, dx = i(2\pi)^{-n/2} \int_{\mathbb{R}^n} e^{it\cdot x} x^{e_j} f(x) \, dx.$$

Now $x^{e_j} f(x) \in \mathcal{S}$ and so one can continue in this way and take derivatives indefinitely. Thus $F^{-1}f \in C^\infty(\mathbb{R}^n)$ and from the above argument,

$$D^\alpha F^{-1}f(t) = (2\pi)^{-n/2} \int_{\mathbb{R}^n} e^{it\cdot x} (ix)^\alpha f(x) dx.$$

To complete showing $F^{-1}f \in \mathcal{S}$,

$$t^\beta D^\alpha F^{-1}f(t) = (2\pi)^{-n/2} \int_{\mathbb{R}^n} e^{it\cdot x} t^\beta (ix)^a f(x) \, dx.$$

Integrate this integral by parts to get

$$t^{\beta}D^{\alpha}F^{-1}f(t)=(2\pi)^{-n/2}\int_{\mathbb{R}^n}i^{|\beta|}e^{it\cdot x}D^{\beta}\left((ix)^{a}f(x)\right)dx. \tag{13.15}$$

Here is how this is done.

$$\begin{aligned}\int_{\mathbb{R}}e^{it_jx_j}t_j^{\beta_j}(ix)^{\alpha}f(x)dx_j &= \frac{e^{it_jx_j}}{it_j}t_j^{\beta_j}(ix)^{\alpha}f(x)\,|_{-\infty}^{\infty}+\\ &\quad i\int_{\mathbb{R}}e^{it_jx_j}t_j^{\beta_j-1}D^{e_j}((ix)^{\alpha}f(x))dx_j\end{aligned}$$

where the boundary term vanishes because $f\in\mathcal{S}$. Returning to 13.15, use the fact that $|e^{ia}|=1$ to conclude

$$\left|t^{\beta}D^{\alpha}F^{-1}f(t)\right|\leq C\int_{\mathbb{R}^n}\left|D^{\beta}\left((ix)^{a}f(x)\right)\right|dx<\infty.$$

It follows $F^{-1}f\in\mathcal{S}$. Similarly $Ff\in\mathcal{S}$ whenever $f\in\mathcal{S}$. ■

Of course $\mathcal{S}$ can be considered a subset of $\mathcal{G}^*$ as follows. For $\psi\in\mathcal{S}$, $\psi(\phi)\equiv\int_{\mathbb{R}^n}\psi\phi dx$.

Theorem 13.2.22 *Let $\psi\in\mathcal{S}$. Then $\left(F\circ F^{-1}\right)(\psi)=\psi$ and $(F^{-1}\circ F)(\psi)=\psi$ whenever $\psi\in\mathcal{S}$. Also F and F^{-1} map $\mathcal{S}$ one to one and onto $\mathcal{S}$.*

Proof: The first claim follows from the fact that F and F^{-1} are inverses of each other on $\mathcal{G}^*$ which was established above. For the second, let $\psi\in\mathcal{S}$. Then $\psi=F\left(F^{-1}\psi\right)$. Thus F maps $\mathcal{S}$ onto $\mathcal{S}$. If $F\psi=0$, then do F^{-1} to both sides to conclude $\psi=0$. Thus F is one to one and onto. Similarly, F^{-1} is one to one and onto. ■

13.2.4 Convolution

To begin with it is necessary to discuss the meaning of ϕf where $f\in\mathcal{G}^*$ and $\phi\in\mathcal{G}$. What should it mean? First suppose $f\in L^p(\mathbb{R}^n)$ or measurable with polynomial growth. Then ϕf also has these properties. Hence, it should be the case that $\phi f(\psi)=\int_{\mathbb{R}^n}\phi f\psi dx=\int_{\mathbb{R}^n}f(\phi\psi)dx$. This motivates the following definition.

Definition 13.2.23 *Let $T\in\mathcal{G}^*$ and let $\phi\in\mathcal{G}$. Then $\phi T\equiv T\phi\in\mathcal{G}^*$ will be defined by*

$$\phi T(\psi)\equiv T(\phi\psi).$$

The next topic is that of convolution. It was just shown that

$$F(f*\phi)=(2\pi)^{n/2}F\phi Ff,\ F^{-1}(f*\phi)=(2\pi)^{n/2}F^{-1}\phi F^{-1}f$$

whenever $f\in L^2(\mathbb{R}^n)$ and $\phi\in\mathcal{G}$ so the same definition is retained in the general case because it makes perfect sense and agrees with the earlier definition.

Definition 13.2.24 *Let $f\in\mathcal{G}^*$ and let $\phi\in\mathcal{G}$. Then define the convolution of f with an element of $\mathcal{G}$ as follows.*

$$f*\phi\equiv(2\pi)^{n/2}F^{-1}(F\phi Ff)\in\mathcal{G}^*$$

There is an obvious question. With this definition, is it true that

$$F^{-1}(f*\phi) = (2\pi)^{n/2} F^{-1}\phi F^{-1} f$$

as it was earlier?

Theorem 13.2.25 *Let $f \in \mathscr{G}^*$ and let $\phi \in \mathscr{G}$.*

$$F(f*\phi) = (2\pi)^{n/2} F\phi F f, \tag{13.16}$$

$$F^{-1}(f*\phi) = (2\pi)^{n/2} F^{-1}\phi F^{-1} f. \tag{13.17}$$

Proof: Note that 13.16 follows from Definition 13.2.24 and both assertions hold for $f \in \mathscr{G}$. Consider 13.17. Here is a simple formula involving a pair of functions in $\mathscr{G}$.

$$\left(\psi * F^{-1}F^{-1}\phi\right)(\boldsymbol{x})$$

$$\begin{aligned} &= \left(\int\int\int \psi(\boldsymbol{x}-\boldsymbol{y})\, e^{i\boldsymbol{y}\cdot\boldsymbol{y}_1} e^{i\boldsymbol{y}_1\cdot\boldsymbol{z}} \phi(\boldsymbol{z})\, dz dy_1 dy\right)(2\pi)^n \\ &= \left(\int\int\int \psi(\boldsymbol{x}-\boldsymbol{y})\, e^{-i\boldsymbol{y}\cdot\tilde{\boldsymbol{y}}_1} e^{-i\tilde{\boldsymbol{y}}_1\cdot\boldsymbol{z}} \phi(\boldsymbol{z})\, dz d\tilde{y}_1 dy\right)(2\pi)^n = (\psi * FF\phi)(\boldsymbol{x}). \end{aligned}$$

Now for $\psi \in \mathscr{G}$,

$$\begin{gathered} (2\pi)^{n/2} F\left(F^{-1}\phi F^{-1} f\right)(\psi) \equiv (2\pi)^{n/2}\left(F^{-1}\phi F^{-1} f\right)(F\psi) \equiv \\ (2\pi)^{n/2} F^{-1} f\left(F^{-1}\phi F\psi\right) \equiv (2\pi)^{n/2} f\left(F^{-1}\left(F^{-1}\phi F\psi\right)\right) = \\ f\left((2\pi)^{n/2} F^{-1}\left(\left(FF^{-1}F^{-1}\phi\right)(F\psi)\right)\right) \equiv \\ f\left(\psi * F^{-1}F^{-1}\phi\right) = f(\psi * FF\phi) \end{gathered} \tag{13.18}$$

Also

$$\begin{gathered} (2\pi)^{n/2} F^{-1}(F\phi F f)(\psi) \equiv (2\pi)^{n/2}(F\phi F f)\left(F^{-1}\psi\right) \equiv \\ (2\pi)^{n/2} F f\left(F\phi F^{-1}\psi\right) \equiv (2\pi)^{n/2} f\left(F\left(F\phi F^{-1}\psi\right)\right) = \\ = f\left(F\left((2\pi)^{n/2}\left(F\phi F^{-1}\psi\right)\right)\right) \\ = f\left(F\left((2\pi)^{n/2}\left(F^{-1}FF\phi F^{-1}\psi\right)\right)\right) = f\left(F\left(F^{-1}\left(FF\phi * \psi\right)\right)\right) \\ f(FF\phi * \psi) = f(\psi * FF\phi). \end{gathered} \tag{13.19}$$

The last line follows from the following.

$$\begin{aligned} \int FF\phi(\boldsymbol{x}-\boldsymbol{y})\,\psi(\boldsymbol{y})\,dy &= \int F\phi(\boldsymbol{x}-\boldsymbol{y})\, F\psi(\boldsymbol{y})\,dy = \int F\psi(\boldsymbol{x}-\boldsymbol{y})\, F\phi(\boldsymbol{y})\,dy \\ &= \int \psi(\boldsymbol{x}-\boldsymbol{y})\, FF\phi(y)\,dy. \end{aligned}$$

From 13.19 and 13.18 , since ψ was arbitrary,

$$(2\pi)^{n/2} F\left(F^{-1}\phi F^{-1} f\right) = (2\pi)^{n/2} F^{-1}(F\phi F f) \equiv f * \phi$$

which shows 13.17. ■

13.3 Exercises

1. For $f \in L^1(\mathbb{R}^n)$, show that if $F^{-1}f \in L^1$ or $Ff \in L^1$, then f equals a continuous bounded function a.e.

2. Suppose $f, g \in L^1(\mathbb{R})$ and $Ff = Fg$. Show $f = g$ a.e.

3. Show that if $f \in L^1(\mathbb{R}^n)$, then $\lim_{|\boldsymbol{x}|\to\infty} Ff(\boldsymbol{x}) = 0$.

4. ↑ Suppose $f * f = f$ or $f * f = 0$ and $f \in L^1(\mathbb{R})$. Show $f = 0$.

5. For this problem define $\int_a^\infty f(t)\,dt \equiv \lim_{r\to\infty} \int_a^r f(t)\,dt$. Note this coincides with the Lebesgue integral when $f \in L^1(a,\infty)$. Show

   (a) $\int_0^\infty \frac{\sin(u)}{u} du = \frac{\pi}{2}$

   (b) $\lim_{r\to\infty} \int_\delta^\infty \frac{\sin(ru)}{u} du = 0$ whenever $\delta > 0$.

   (c) If $f \in L^1(\mathbb{R})$, then $\lim_{r\to\infty} \int_{\mathbb{R}} \sin(ru) f(u)\,du = 0$.

   Hint: For the first two, use $\frac{1}{u} = \int_0^\infty e^{-ut}dt$ and then, using this, apply Fubini's theorem to $\int_0^R \sin u \int_{\mathbb{R}} e^{-ut}\,dt\,du$. For the last part, first establish it for $f \in C_c^\infty(\mathbb{R})$ and then use the density of this set in $L^1(\mathbb{R})$ to obtain the result. This is sometimes called the Riemann Lebesgue lemma.

6. ↑Suppose that $g \in L^1(\mathbb{R})$ and that at some $x > 0$, g is locally Holder continuous from the right and from the left. This means

$$\lim_{r\to 0+} g(x+r) \equiv g(x+), \ \lim_{r\to 0+} g(x-r) \equiv g(x-)$$

   exists and there exist constants $K, \delta > 0$ and $r \in (0,1]$ such that for $|x-y| < \delta$,

$$|g(x+) - g(y)| < K|x-y|^r, \ |g(x-) - g(y)| < K|x-y|^r$$

   for $y > x$ and $y < x$ respectively. Show that under these conditions,

$$\lim_{r\to\infty} \frac{2}{\pi} \int_0^\infty \frac{\sin(ur)}{u} \left(\frac{g(x-u) + g(x+u)}{2} \right) du = \frac{g(x+) + g(x-)}{2}.$$

7. Let $g \in L^1(\mathbb{R})$ and suppose g is locally Holder continuous from the right and from the left at x. Show that then

$$\lim_{R\to\infty} \frac{1}{2\pi} \int_{-R}^R e^{ixt} \int_{-\infty}^\infty e^{-ity} g(y)\,dy\,dt = \frac{g(x+) + g(x-)}{2}.$$

   This is very interesting. If $g \in L^2(\mathbb{R})$, this shows $F^{-1}(Fg)(x) = \frac{g(x+)+g(x-)}{2}$, the midpoint of the jump in g at the point, x. In particular, if $g \in \mathcal{G}$, $F^{-1}(Fg) = g$. **Hint:** Show the left side of the above equation reduces to

$$\frac{2}{\pi} \int_0^\infty \frac{\sin(ur)}{u} \left(\frac{g(x-u) + g(x+u)}{2} \right) du$$

   and then use Problem 6 to obtain the result.

8. ↑ A measurable function g defined on $(0,\infty)$ has exponential growth if $|g(t)| \leq Ce^{\eta t}$ for some η. For $\mathrm{Re}\,(s) > \eta$, define the Laplace Transform: $Lg(s) \equiv \int_0^\infty e^{-su} g(u)\,du$. Assume that g has exponential growth as above and is Holder continuous from the right and from the left at t. Pick $\gamma > \eta$. Show that

$$\lim_{R\to\infty} \frac{1}{2\pi} \int_{-R}^{R} e^{\gamma t} e^{iyt} Lg(\gamma + iy)\,dy = \frac{g(t+) + g(t-)}{2}.$$

This formula is sometimes written in the form $\frac{1}{2\pi i}\int_{\gamma-i\infty}^{\gamma+i\infty} e^{st} Lg(s)\,ds$ and is called the complex inversion integral for Laplace transforms. It can be used to find inverse Laplace transforms. **Hint:** Plug in the formula for the Laplace transform and then massage to get it in the form of the preceding problem.

9. Suppose $f \in \mathscr{G}$. Show $F(f_{x_j})(\boldsymbol{t}) = it_j Ff(\boldsymbol{t})$.

10. Let $f \in \mathscr{G}$ and let k be a positive integer. $\|f\|_{k,2} \equiv (\|f\|_2^2 + \sum_{|\alpha|\leq k} \|D^\alpha f\|_2^2)^{1/2}$. One could also define $\|f\|'_{k,2} \equiv (\int_{R^n} |Ff(\boldsymbol{x})|^2 (1+|\boldsymbol{x}|^2)^k dx)^{1/2}$. Show both $\|\cdot\|_{k,2}$ and $\|\cdot\|'_{k,2}$ are norms on $\mathscr{G}$ and that they are equivalent. These are Sobolev space norms. For which values of k does the second norm make sense? How about the first norm?

11. ↑Define $H^k(\mathbb{R}^n), k \geq 0$ by $f \in L^2(\mathbb{R}^n)$ such that

$$\left(\int |Ff(\boldsymbol{x})|^2 (1+|\boldsymbol{x}|^2)^k dx\right)^{\frac{1}{2}} < \infty,\ \|f\|'_{k,2} \equiv \left(\int |Ff(\boldsymbol{x})|^2 (1+|\boldsymbol{x}|^2)^k dx\right)^{\frac{1}{2}}.$$

Show $H^k(\mathbb{R}^n)$ is a Banach space, and that if k is a positive integer, $H^k(\mathbb{R}^n)$ will be the set of all $f \in L^2(\mathbb{R}^n)$ such that there exists $\{u_j\} \subseteq \mathscr{G}$ with $\|u_j - f\|_2 \to 0$ and $\{u_j\}$ is a Cauchy sequence in $\|\ \|_{k,2}$ of Problem 10. This is one way to define Sobolev Spaces. **Hint:** One way to do the second part of this is to define a new measure μ by $\mu(E) \equiv \int_E \left(1+|\boldsymbol{x}|^2\right)^k dx$. Then show μ is a Borel measure which is inner and outer regular and show there exists $\{g_m\}$ such that $g_m \in \mathscr{G}$ and $g_m \to Ff$ in $L^2(\mu)$. Thus $g_m = Ff_m$, $f_m \in \mathscr{G}$ because F maps $\mathscr{G}$ onto $\mathscr{G}$. Then by Problem 10, $\{f_m\}$ is Cauchy in the norm $\|\ \|_{k,2}$.

12. ↑ If $2k > n$, show that if $f \in H^k(\mathbb{R}^n)$, then f equals a bounded continuous function a.e. **Hint:** Show that for k this large, $Ff \in L^1(\mathbb{R}^n)$, and then use Problem 1. To do this, write

$$|Ff(\boldsymbol{x})| = |Ff(\boldsymbol{x})|(1+|\boldsymbol{x}|^2)^{\frac{k}{2}}(1+|\boldsymbol{x}|^2)^{\frac{-k}{2}},$$

So $\int |Ff(\boldsymbol{x})|dx = \int |Ff(\boldsymbol{x})|(1+|\boldsymbol{x}|^2)^{\frac{k}{2}}(1+|\boldsymbol{x}|^2)^{\frac{-k}{2}} dx$. Use the Cauchy Schwarz inequality. This is an example of a Sobolev imbedding Theorem.

13. Let $u \in \mathscr{G}$. Then $Fu \in \mathscr{G}$ and so, in particular, it makes sense to form the integral,

$$\int_{\mathbb{R}} Fu\left(\boldsymbol{x}', x_n\right) dx_n$$

where $(\boldsymbol{x}', x_n) = \boldsymbol{x} \in \mathbb{R}^n$. For $u \in \mathscr{G}$, define $\gamma u(\boldsymbol{x}') \equiv u(\boldsymbol{x}', 0)$. Find a constant such that $F(\gamma u)(\boldsymbol{x}')$ equals this constant times the above integral. **Hint:** By the dominated convergence theorem

$$\int_{\mathbb{R}} Fu\left(\boldsymbol{x}', x_n\right) dx_n = \lim_{\varepsilon\to 0} \int_{\mathbb{R}} e^{-(\varepsilon x_n)^2} Fu\left(\boldsymbol{x}', x_n\right) dx_n.$$

Now use the definition of the Fourier transform and Fubini's theorem as required in order to obtain the desired relationship.

14. Let $h(x) = \left(\int_0^x e^{-t^2}dt\right)^2 + \left(\int_0^1 \frac{e^{-x^2\left(1+t^2\right)}}{1+t^2}dt\right)$. Show that $h'(x) = 0$ and $h(0) = \pi/4$. Then let $x \to \infty$ to conclude that $\int_0^\infty e^{-t^2}dt = \sqrt{\pi}/2$. Show that $\int_{-\infty}^\infty e^{-t^2}dt = \sqrt{\pi}$ and that $\int_{-\infty}^\infty e^{-ct^2}dt = \frac{\sqrt{\pi}}{\sqrt{c}}$.

15. Recall that for f a function, $f_{\boldsymbol{y}}(\boldsymbol{x}) = f(\boldsymbol{x}-\boldsymbol{y})$. Find a relationship between $Ff_{\boldsymbol{y}}(\boldsymbol{t})$ and $Ff(\boldsymbol{t})$ given that $f \in L^1(\mathbb{R}^n)$.

16. For $f \in L^1(\mathbb{R}^n)$, simplify $Ff(\boldsymbol{t}+\boldsymbol{y})$.

17. For $f \in L^1(\mathbb{R}^n)$ and c a nonzero real number, show $Ff(c\boldsymbol{t}) = Fg(\boldsymbol{t})$ where $g(\boldsymbol{x}) = f\left(\frac{\boldsymbol{x}}{c}\right)$.

18. Suppose that $f \in L^1(\mathbb{R})$ and that $\int |x|\,|f(x)|\,dx < \infty$. Find a way to use the Fourier transform of f to compute $\int xf(x)\,dx$.

19. Let $(\Omega, \mathscr{F}, P)$ be a probability space and let $X : \Omega \to \mathbb{R}^n$ be a random variable. This means $X^{-1}(\text{open set}) \in \mathscr{F}$. Define a measure λ_X on the Borel sets of $\mathbb{R}^n$ as follows. For E a Borel set, $\lambda_X(E) \equiv P\left(X^{-1}(E)\right)$. Explain why this is well defined. Next explain why λ_X can be considered a Radon probability measure by completion. Explain why $\lambda_X \in \mathscr{G}^*$ if $\lambda_X(\psi) \equiv \int_{\mathbb{R}^n} \psi d\lambda_X$ where $\mathscr{G}$ is the collection of functions used to define the Fourier transform.

20. Using the above problem, the characteristic function of this measure (random variable) is $\phi_X(\boldsymbol{y}) \equiv \int_{\mathbb{R}^n} e^{i\boldsymbol{x}\cdot\boldsymbol{y}}d\lambda_X$. Show this always exists for any such random variable and is continuous. Next show that for two random variables $X, Y, \lambda_X = \lambda_Y$ if and only if $\phi_X(\boldsymbol{y}) = \phi_Y(\boldsymbol{y})$ for all $\boldsymbol{y}$. In other words, show the distribution measures are the same if and only if the characteristic functions are the same. A lot more can be concluded by looking at characteristic functions of this sort. The important thing about these characteristic functions is that they always exist, unlike moment generating functions.

Chapter 14

Integration on Manifolds

Till now, integrals have mostly pertained to measurable subsets of $\mathbb{R}^p$ and not something like a surface contained in a higher dimensional space. This is what is considered in this chapter. First is an abstract description of manifolds and then an interesting application of the representation theorem for positive linear functionals is used to give a measure on a manifold. This is the higher dimensional version of arc length for a smooth curve seen in calculus. This section can be omitted if desired except for the material on the divergence theorem because it will be done much better later.

14.1 Relatively Open Sets

An essential part of the definition of a manifold is the idea of a relatively open set defined next.

Definition 14.1.1 *Let X be a normed linear space and let $\Omega \subseteq X$. Then a set U is called a relatively open set or open in Ω if it is the intersection of an open set of X with Ω. Thus Ω is a metric space with respect to the distance $d(x,y) \equiv \|x-y\|$ and all considerations such as limit points, closures, limits, closed sets, open sets etc. in this metric space are taken with respect to this metric. Continuity is also defined in terms of this metric on Ω inherited from X. If U is open in Ω and $f : U \to Y$ where Y is a normed linear space, then f is differentiable on U means that there exists $\hat{U}$ open in X such that $f = g$ on U and $g : \hat{U} \to Y$ is differentiable on $\hat{U}$. In words, f is the restriction to U of a differentiable function defined on an open set of X.*

14.2 Manifolds

Manifolds are sets which resemble an open subset of $\mathbb{R}^p$ locally. A manifold with boundary resembles locally a set open in $\mathbb{R}^p_+$. One needs the theorem on invariance of domain, Theorem 8.10.5 to verify the boundary is well defined. There are many other ways to generalize what is presented here. For example, it will be assumed here that the atlas is finite. It suffices to assume it is locally finite. It is not necessary to assume the manifold is a compact subset of $\mathbb{R}^q$ as is done below and so forth. It is enough to have it a metric space. To make this concept of a manifold more precise, here is a definition.

Definition 14.2.1 *Let $V \subseteq \mathbb{R}^q$. $C^k\left(\overline{V};\mathbb{R}^p\right)$ is the set of functions which are restrictions to V of some function defined on $\mathbb{R}^q$ which has k continuous derivatives which has values in $\mathbb{R}^p$. When $k = 0$, it means the restriction to V of continuous functions. A function is in $D\left(\overline{V};\mathbb{R}^p\right)$ if it is the restriction to V of a differentiable function defined on $\mathbb{R}^q$. A Lipschitz function $\boldsymbol{f}$ is one which satisfies $\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{y})\| \leq K\|\boldsymbol{x}-\boldsymbol{y}\|$.*

Thus, if $f \in C^k\left(\overline{V};\mathbb{R}^q\right)$ or $D\left(\overline{V};\mathbb{R}^p\right)$, we can consider it defined on $\overline{V}$ and not just on V. This is the way one can generalize a one sided derivative of a function defined on a closed interval. First consider the case of a topological manifold in which there is no differentiability involved.

Definition 14.2.2 *A metric space Ω, will be called a p dimensional C^0 manifold, $p \geq 1$, if for each $\boldsymbol{x} \in \Omega$, there exists an open set $U_{\boldsymbol{x}}$ containing $\boldsymbol{x}$ and a one to one continuous mapping $\boldsymbol{R}_{\boldsymbol{x}} : U_{\boldsymbol{x}} \to \mathbb{R}^p_+$ whose inverse is also continuous. Such a pair $(U, \boldsymbol{R})$*

is called a chart. A point $\boldsymbol{x}$ on Ω is called a boundary point of Ω, indicated by $\boldsymbol{x} \in \partial\Omega$, means that for $(U, \boldsymbol{R})$ a chart for which $\boldsymbol{x} \in U$, $\boldsymbol{R}\boldsymbol{x} \in \mathbb{R}^p_0 \equiv \{\boldsymbol{x} \in \mathbb{R}^p : x_p = 0\}$.

First of all I need to show that the concept of $\partial\Omega$ is well defined.

Lemma 14.2.3 *$\partial\Omega$ is well defined in the sense that the statement that $\boldsymbol{x}$ is a boundary point does not depend on which chart is considered.*

Proof: Suppose $\boldsymbol{x}$ is not a boundary point with respect to the chart $(U, \boldsymbol{R})$ but is a boundary point with respect to $(V, \boldsymbol{S})$. Then $(\boldsymbol{R}\boldsymbol{x})_p > 0$ so there is an open ball $B_{\boldsymbol{x}}$ in $\mathbb{R}^p$centered at $\boldsymbol{x}$, $B_{\boldsymbol{x}} \subseteq \boldsymbol{R}(V \cap U)$. By invariance of domain theorem,Theorem 8.10.5, $\boldsymbol{S} \circ \boldsymbol{R}^{-1}(B_{\boldsymbol{x}})$ is a set open in $\mathbb{R}^p$ which is contained in $\boldsymbol{S}(V \cap U)$ which violates the claim that $\boldsymbol{x}$ is a boundary point with respect to $(V, \boldsymbol{S})$. This shows that the definition of boundary points is well defined. ■

I will assume in what follows that Ω is a compact subset of $\mathbb{R}^q$, $q \geq p$ and is a metric space coming from the Euclidean norm $|\cdot|$ on Ω. You could get by with replacing $\mathbb{R}^q$ with a Hilbert space as in Problem 9 on Page 152 in place of $\mathbb{R}^q$.

Definition 14.2.4 *A compact subset Ω of $\mathbb{R}^q$ will be called a differentiable p dimensional manifold with boundary if it is a C^0 manifold and also has some differentiable structure about to be described. By compactness, there are finitely many charts $\{(U_i, \boldsymbol{R}_i)\}$ (described above) such that the U_i cover Ω. This set of charts is called an atlas. Ω is differentiable if it has an atlas such that $\boldsymbol{R}_j$ and $\boldsymbol{R}_j^{-1}$ are differentiable on U_j and $\boldsymbol{R}_j(U_j)$ respectively. If, in addition to this, it has an atlas $(U_i, \boldsymbol{R}_i)$ such that*

$$\det\left(D\boldsymbol{R}_i^{-1}(\boldsymbol{R}_i(\boldsymbol{x}))\right)^* \left(D\boldsymbol{R}_i^{-1}(\boldsymbol{R}_i(\boldsymbol{x}))\right) \neq 0$$

then it is a smooth manifold. This condition is like the one for a smooth curve in calculus in which the derivative does not vanish. If "differentiable" is replaced with C^k meaning the first k derivatives exist and are continuous, then it will be a smooth C^k manifold with boundary.

Definition 14.2.5 *A C^0 manifold is called a Lipschitz manifold if all the maps and their inverses in a typical chart are Lipschitz. Thus $\boldsymbol{S}_j \circ \boldsymbol{R}_i^{-1}$ is Lipschitz if $\{(\boldsymbol{S}_j, V_j)\}$ is another such chart because the composition of Lipschitz functions is Lipschitz.*

It turns out that Lipschitz functions are differentiable a.e. This is Rademacher's theorem presented later. This situation is easier to consider. It is not clear that a Lipschitz function is the restriction of an everywhere differentiable function to the complement of a set of measure zero although it will be shown later that a function Lipschitz on a subset of $\mathbb{R}^n$ is the restriction of a Lipschitz function defined on $\mathbb{R}^n$. With Rademacher's theorem, everything presented here goes through with no changes when applied to Lipschitz mappings.

Lemma 14.2.6 *Let Ω be a differentiable manifold and let $\{(\boldsymbol{R}_i, U_i)\}, \{(\boldsymbol{S}_j, V_j)\}$ both be an atlas as described above. Then $\boldsymbol{S}_j \circ \boldsymbol{R}_i^{-1}$ is differentiable on $\boldsymbol{R}_i(U_i \cap V_j)$.*

Proof: This follows from the chain rule and the consideration of the definition that $\boldsymbol{S}_j$ is the restriction of a differentiable function. ■

Next is the concept of an oriented manifold. Orientation can be defined for general C^0 manifolds using the topological degree, but the reason for considering this, at least here, involves some sort of differentiability.

Definition 14.2.7 *A differentiable manifold Ω with boundary is called orientable if there exists an atlas, $\{(U_r, \boldsymbol{R}_r)\}_{r=1}^{p}$, such that whenever $U_i \cap U_j \neq \emptyset$,*

$$\det\left(D\left(\boldsymbol{R}_j \circ \boldsymbol{R}_i^{-1}\right)\right)(\boldsymbol{u}) > 0 \text{ for all } \boldsymbol{u} \in \boldsymbol{R}_i(U_i \cap U_j) \tag{14.1}$$

This makes sense from Lemma 14.2.6. An atlas satisfying 14.1 is called an oriented atlas. Also the following notation is often used with the convention that $\boldsymbol{v} = \boldsymbol{R}_i \circ \boldsymbol{R}_j^{-1}(\boldsymbol{u})$

$$\frac{\partial(v_1 \cdots v_p)}{\partial(u_1 \cdots u_p)} \equiv \det D\left(\boldsymbol{R}_i \circ \boldsymbol{R}_j^{-1}\right)(\boldsymbol{u})$$

In this case, another atlas will be called an equivalent atlas $(V_i, \boldsymbol{S}_i)$ if

$$\det\left(D\left(\boldsymbol{S}_j \circ \boldsymbol{R}_i^{-1}\right)\right)(\boldsymbol{u}) > 0 \text{ for all } \boldsymbol{u} \in \boldsymbol{R}_i(U_i \cap V_j)$$

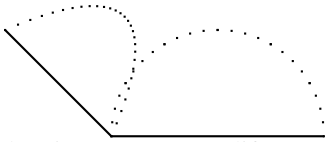
You can verify that this condition does indeed define an equivalence relation. Thus an oriented atlas would consist of a set of points along with an equivalence class of atlases. You could also define a piecewise differentiable manifold as the union of finitely many differentiable manifolds which have intersection only at boundary points.

This is like a piecewise differentiable curve from calculus.

Example 14.2.8 *Let $f : \mathbb{R}^{p+1} \to \mathbb{R}$ is C^1 and suppose and that $Df(\boldsymbol{x}) \neq \boldsymbol{0}$ for all $\boldsymbol{x}$ contained in the set $\{\boldsymbol{x} : f(\boldsymbol{x}) = 0\}$. Then if $\{\boldsymbol{x} : f(\boldsymbol{x}) = 0\}$ is nonempty, it is a C^1 manifold thanks to an application of the implicit function theorem.*

Note that this includes $S^{p-1}, \{\boldsymbol{x} \in \mathbb{R}^p : |\boldsymbol{x}| = 1\}$ and lots of other things like $x^4 + y^2 + z^4 = 1$ and so forth. The details are left as an exercise.

What if the boundary of Ω has pointy places so it is not differentiable. Does this prevent Ω from being a differentiable manifold? Not necessarily. Consider the following picture where there is a pointy place on the boundary of Ω. Then in this case, there is still a way to obtain the necessary kind of chart. See the following picture.



These kinds of fussy considerations seem to proliferate. The solution is to use Hausdorff measures and Rademacher's theorem to simultaneously simplify and generalize. This is presented later. One can also emphasize parametrizations of manifolds instead of fussing with charts. The idea of a topological manifold is very elegant but when you start trying to do calculus on them, some of the elegence is lost in a forest of minutia. With Rademacher's theorem, it isn't so bad. Nevertheless, the next section will consider integration on differentiable manifolds as an application of the representation theorem for positive linear functionals.

14.3 The Area Measure on a Manifold

Next the "surface measure" on a manifold is given. In what follows, the manifold will be a compact subset of $\mathbb{R}^q$. This has nothing to do with orientation. It will involve the following

definition. To motivate this definition, recall the way you found the length of a curve in calculus where $t \in [a,b]$. It was

$$\int_a^b \left|\boldsymbol{r}'(t)\right| dt = \int_a^b \det\left(D\boldsymbol{r}(t)^* D\boldsymbol{r}(t)\right)^{1/2} dt.$$

where $\boldsymbol{r}(t)$ is a parametrization for the curve. Think of $dl = \det\left(D\boldsymbol{r}(t)^* D\boldsymbol{r}(t)\right)^{1/2} dt$ and you sum these to get the length.

Definition 14.3.1 *Let $(U_i, \boldsymbol{R}_i)$ be an atlas for a p dimensional differentiable manifold with boundary Ω. Also let $\{\psi_i\}_{i=1}^r$ be a partition of unity from Theorem 3.12.5 $\operatorname{spt}\psi_i \subseteq U_i$. Then for $f \in C_c(\Omega)$, define*

$$Lf \equiv \sum_{i=1}^r \int_{\boldsymbol{R}_i(U_i)} f\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right) \psi_i\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right) J_i(\boldsymbol{u})\, du$$

Here du signifies $dm_p(\boldsymbol{u})$ and

$$J_i(\boldsymbol{u}) \equiv \left(\det\left(D\boldsymbol{R}_i^{-1}(\boldsymbol{u})^* D\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right)\right)^{1/2}$$

I need to show that the same thing is obtained if another atlas and/or partition of unity is used.

Theorem 14.3.2 *The functional L is well defined in the sense that if another atlas is used, then for $f \in C_c(\Omega)$, the same value is obtained for Lf.*

Proof: Let the other atlas be $\left\{(V_j, \boldsymbol{S}_j)\right\}_{j=1}^s$ where $\boldsymbol{v} \in V_j$ and $\boldsymbol{S}_j$ has the same properties as the $\boldsymbol{R}_i$. Then $\left(\boldsymbol{S}_j \circ \boldsymbol{R}_i^{-1}\right)(\boldsymbol{u}) = \boldsymbol{v}$ so $\boldsymbol{R}_i^{-1}(\boldsymbol{u}) = \boldsymbol{S}_j^{-1}(\boldsymbol{v})$ and so $\boldsymbol{R}_i^{-1}(\boldsymbol{u}) = \boldsymbol{S}_j^{-1}\left(\left(\boldsymbol{S}_j \circ \boldsymbol{R}_i^{-1}\right)(\boldsymbol{u})\right)$ implying $D\boldsymbol{R}_i^{-1}(\boldsymbol{u}) = D\boldsymbol{S}_j^{-1}(\boldsymbol{v}) D\left(\boldsymbol{S}_j \circ \boldsymbol{R}_i^{-1}\right)(\boldsymbol{u})$. Therefore,

$$J_i(\boldsymbol{u}) = \left(\det\left(D\boldsymbol{R}_i^{-1}(\boldsymbol{u})^* D\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right)\right)^{1/2}$$

$$= \left(\det\left(\overbrace{D\left(\boldsymbol{S}_j \circ \boldsymbol{R}_i^{-1}\right)^*(\boldsymbol{u})}^{p\times p} \overbrace{D\boldsymbol{S}_j^{-1}(\boldsymbol{v})^*}^{(p\times q)} \overbrace{D\boldsymbol{S}_j^{-1}(\boldsymbol{v}) D\left(\boldsymbol{S}_j \circ \boldsymbol{R}_i^{-1}\right)(\boldsymbol{u})}^{(q\times p)\quad p\times p}\right)\right)^{1/2}$$

$$= \left[\det\left(D\left(\boldsymbol{S}_j \circ \boldsymbol{R}_i^{-1}\right)^*(\boldsymbol{u})\right) \det\left(D\left(\boldsymbol{S}_j \circ \boldsymbol{R}_i^{-1}\right)(\boldsymbol{u})\right)\right]^{1/2} J_j(\boldsymbol{v})$$

$$= \left|\det\left(D\left(\boldsymbol{S}_j \circ \boldsymbol{R}_i^{-1}\right)(\boldsymbol{u})\right)\right| J_j(\boldsymbol{v}) \tag{14.2}$$

Similarly

$$J_j(\boldsymbol{v}) = \left|\det\left(D\left(\boldsymbol{R}_i \circ \boldsymbol{S}_j^{-1}\right)(\boldsymbol{v})\right)\right| J_i(\boldsymbol{u}). \tag{14.3}$$

Let $\hat{L}$ go with this new atlas. Thus

$$\hat{L}(f) \equiv \sum_{j=1}^s \int_{\boldsymbol{S}_j(V_j)} f\left(\boldsymbol{S}_j^{-1}(\boldsymbol{v})\right) \eta_j\left(\boldsymbol{S}_j^{-1}(\boldsymbol{v})\right) J_j(\boldsymbol{v})\, dv \tag{14.4}$$

where η_j is a partition of unity associated with the sets V_j as described above. Now letting ψ_i be the partition of unity for the U_i, $\boldsymbol{v} = \boldsymbol{S}_j \circ \boldsymbol{R}_i^{-1}(\boldsymbol{u})$ for $\boldsymbol{u} \in \boldsymbol{R}_i(V_j \cap U_i)$.

$$\begin{aligned} & \int_{\boldsymbol{S}_j(V_j)} f\left(\boldsymbol{S}_j^{-1}(\boldsymbol{v})\right) \eta_j\left(\boldsymbol{S}_j^{-1}(\boldsymbol{v})\right) J_j(\boldsymbol{v})\, dv \\ = & \sum_{i=1}^{r} \int_{\boldsymbol{S}_j(V_j \cap U_i)} f\left(\boldsymbol{S}_j^{-1}(\boldsymbol{v})\right) \psi_i\left(\boldsymbol{S}_j^{-1}(\boldsymbol{v})\right) \eta_j\left(\boldsymbol{S}_j^{-1}(\boldsymbol{v})\right) J_j(\boldsymbol{v})\, dv \end{aligned}$$

By Lemma 11.14.1, the assumptions of differentiability imply that the boundary points of Ω are always mapped to a set of measure zero so these can be neglected if desired. Now $\boldsymbol{S}_j(V_j \cap U_i) = \boldsymbol{S}_j \circ \boldsymbol{R}_i^{-1}(\boldsymbol{R}_i(V_j \cap U_i))$ and so using 14.3, the above expression equals

$$\begin{aligned} & \sum_{i=1}^{r} \int_{\boldsymbol{R}_i(V_j \cap U_i)} f\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right) \psi_i\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right) \eta_j\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right) \cdot \\ & \left|\det\left(D\left(\boldsymbol{R}_i \circ \boldsymbol{S}_j^{-1}\right)(\boldsymbol{v})\right)\right| J_i(\boldsymbol{u}) \left|\det D\left(\boldsymbol{S}_j \circ \boldsymbol{R}_i^{-1}\right)(\boldsymbol{u})\right| du \end{aligned}$$

Now $I = \left(\boldsymbol{R}_i \circ \boldsymbol{S}_j^{-1}\right) \circ \left(\boldsymbol{S}_j \circ \boldsymbol{R}_i^{-1}\right)$ and so the chain rule implies that the product of the two Jacobians is 1. Hence 14.4 equals

$$\begin{aligned} & \sum_{j=1}^{s} \sum_{i=1}^{r} \int_{\boldsymbol{R}_i(V_j \cap U_i)} f\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right) \psi_i\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right) \eta_j\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right) J_i(\boldsymbol{u})\, du \\ = & \sum_{i=1}^{r} \sum_{j=1}^{s} \int_{\boldsymbol{R}_i(U_i)} f\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right) \psi_i\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right) \eta_j\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right) J_i(\boldsymbol{u})\, du \\ = & \sum_{i=1}^{r} \int_{\boldsymbol{R}_i(U_i)} f\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right) \psi_i\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right) \sum_{j=1}^{s} \eta_j\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right) J_i(\boldsymbol{u})\, du \\ = & \sum_{i=1}^{r} \int_{\boldsymbol{R}_i(U_i)} f\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right) \psi_i\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right) J_i(\boldsymbol{u})\, du = L(f) \end{aligned}$$

Thus L is a well defined positive linear functional. ■

Definition 14.3.3 *By the representation theorem for positive linear functionals, Theorem 11.9.2, there exists a complete Radon measure σ_p defined on the Borel sets of Ω such that $Lf = \int_\Omega f d\sigma_p$. Then σ_p is what is meant by the measure on the differentiable manifold Ω.*

If O is an open set in Ω, what is $\sigma_p(O)$? Let $f_n \uparrow \mathcal{X}_O$ where f_n is continuous. Then by the monotone convergence theorem,

$$\begin{aligned} \sigma_p(O) &= \lim_{n\to\infty} L(f_n) = \lim_{n\to\infty} \sum_{i=1}^{r} \int_{\boldsymbol{R}_i(U_i)} f_n\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right) \psi_i\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right) J_i(\boldsymbol{u})\, du \\ &= \lim_{n\to\infty} \sum_{i=1}^{r} \int_{\boldsymbol{R}_i(U_i \cap O)} f_n\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right) \psi_i\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right) J_i(\boldsymbol{u})\, du \\ &= \sum_{i=1}^{r} \int_{\boldsymbol{R}_i(U_i \cap O)} \mathcal{X}_O\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right) \psi_i\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right) J_i(\boldsymbol{u})\, du. \end{aligned}$$

If K is a compact subset of some U_i, then use Corollary 12.6.5 to obtain a partition of unity which has $\psi_i = 1$ on K so that all other ψ_j equal 0. Then

$$\int_\Omega \mathscr{X}_K d\sigma_p = \int_{\boldsymbol{R}_i(U_i)} \mathscr{X}_K\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right) J_i(\boldsymbol{u})\, du$$

It then follows from regularity of the measure and the monotone convergence theorem that if E is any measurable set contained in U_i, you can replace K in the above with E. In general, this implies that for nonnegative measurable f, having support in U_i,

$$\int_\Omega f d\sigma_p = \int_{\boldsymbol{R}_i(U_i)} f\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right) J_i(\boldsymbol{u})\, du$$

Indeed, $\partial\Omega$ is a closed subset of Ω and so $\mathscr{X}_{\partial\Omega}$ is measurable. That part of the boundary contained in U_i would then involve a Lebesgue integral over a part of $\mathbb{R}_0^p$ which is a set of measure zero. This shows the following proposition.

Proposition 14.3.4 *Let Ω be a differentiable manifold as discussed above and let σ_p be the measure on the manifold defined above. Then $\sigma_p(\partial\Omega) = 0$.*

Using Rademacher's theorem which is presented later, which says that every Lipschitz function defined on an open set is differentiable a.e. and Theorem 11.13.1 which says that Lipschitz functions map sets of measure zero to sets of measure zero and measurable sets to measurable sets, the following corollary is obtained using the same arguments.

Corollary 14.3.5 *Let Ω be a subset of a finite dimensional normed linear space be a Lipschitz manifold meaning that it is a C^0 manifold for which each atlas $(U_i, \boldsymbol{R}_i)$ has $\boldsymbol{R}_i$ Lipchitz on U_i and $\boldsymbol{R}_i^{-1}$ is Lipschitz on $\boldsymbol{R}_i(U_i)$. Then there is a regular complete measure σ_p defined on a σ algebra $\mathscr{F}$ of subsets of Ω which is finite on compact sets, includes the Borel sets, and satisfies*

$$\int_\Omega f d\sigma_p = \sum_{i=1}^r \int_{\boldsymbol{R}_i(U_i)} f\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right) \psi_i\left(\boldsymbol{R}_i^{-1}(\boldsymbol{u})\right) J_i(\boldsymbol{u})\, du$$

for all $f \in C_c(\Omega)$ where here the partition of unity comes from Theorem 3.12.5 with respect to the distance coming from the norm.

The justification for using $J_i(\boldsymbol{u})\, du$ in the definition of the area measure comes from geometric reasoning which has not been presented yet. This will be done in the chapter on Hausdorff measures. However, it was noted above that this is a generalization of a familiar example from calculus. It would also be possible to verify that it works from familiar definitions in calculus in the case of a two dimensional manifold. Also note that it suffices to assume only that $D\boldsymbol{R}_i^{-1}(\boldsymbol{u})$ exists for a.e. $\boldsymbol{u}$.

14.4 Divergence Theorem

The divergence theorem considered here will feature an open set in $\mathbb{R}^p$ whose boundary has a particular form. For convenience, if $\boldsymbol{x} \in \mathbb{R}^p$, $\hat{\boldsymbol{x}}_i \equiv \begin{pmatrix} x_1 & \cdots & x_{i-1} & x_{i+1} & \cdots & x_p \end{pmatrix}^T$.

Definition 14.4.1 *Let $U \subseteq \mathbb{R}^p$ satisfy the following conditions. There exist open boxes, $Q_1, \cdots, Q_N$,*

$$Q_i = \prod_{j=1}^{p} (a_j^i, b_j^i)$$

such that $\partial U \equiv \overline{U} \setminus U$ is contained in their union. Also, there exists an open set, Q_0 such that $Q_0 \subseteq \overline{Q_0} \subseteq U$ and $\overline{U} \subseteq Q_0 \cup Q_1 \cup \cdots \cup Q_N$. Assume for each Q_i, there exists k and a function g_i such that $U \cap Q_i$ is of the form

$$\left\{ \begin{array}{c} \boldsymbol{x} : (x_1, \cdots, x_{k-1}, x_{k+1}, \cdots, x_p) \in \prod_{j=1}^{k-1} \left(a_j^i, b_j^i\right) \times \\ \prod_{j=k+1}^{p} \left(a_j^i, b_j^i\right) \text{ and } a_k^i < x_k < g_i(x_1, \cdots, x_{k-1}, x_{k+1}, \cdots, x_p) \end{array} \right\} \tag{14.5}$$

or else of the form

$$\left\{ \begin{array}{c} \boldsymbol{x} : (x_1, \cdots, x_{k-1}, x_{k+1}, \cdots, x_p) \in \prod_{j=1}^{k-1} \left(a_j^i, b_j^i\right) \times \\ \prod_{j=k+1}^{p} \left(a_j^i, b_j^i\right) \text{ and } g_i(x_1, \cdots, x_{k-1}, x_{k+1}, \cdots, x_p) < x_k < b_j^i \end{array} \right\} \tag{14.6}$$

The function, g_i is differentiable and has a measurable partial derivatives on

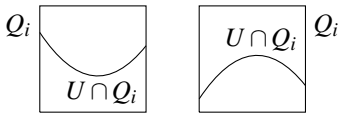
$$A_i \subseteq \prod_{j=1}^{k-1} (a_j^i, b_j^i) \times \prod_{j=k+1}^{p} (a_j^i, b_j^i) \equiv \hat{Q}_k$$

where

$$m_{p-1}\left(\prod_{j=1}^{k-1} (a_j^i, b_j^i) \times \prod_{j=k+1}^{p} (a_j^i, b_j^i) \setminus A_i \right) = 0.$$

and we assume there is a constant C such that for all i and j, $\left| \frac{\partial g_i}{\partial x_j} \right| \leq C$ and that in each variable, g can be recovered from integrating an appropriate partial derivative. That is, each g_k is absolutely continuous in each variable.

To illustrate the above here is a picture.



Recall from calculus that if $z - g(\hat{\boldsymbol{x}}) = 0$ then to get a normal vector to the level surface, it will be $\pm$ the gradient.

Lemma 14.4.2 *Let $\alpha_1, \cdots, \alpha_p$ be real numbers and let $A(\alpha_1, \cdots, \alpha_p)$ be the matrix which has $1 + \alpha_i^2$ in the ii^{th} slot and $\alpha_i \alpha_j$ in the ij^{th} slot when $i \neq j$. Then* $\det A = 1 + \sum_{i=1}^{p} \alpha_i^2$.

Proof of the claim: The matrix, $A(\alpha_1, \cdots, \alpha_p)$ is of the form

$$A(\alpha_1, \cdots, \alpha_p) = \begin{pmatrix} 1+\alpha_1^2 & \alpha_1\alpha_2 & \cdots & \alpha_1\alpha_p \\ \alpha_1\alpha_2 & 1+\alpha_2^2 & & \alpha_2\alpha_p \\ \vdots & & \ddots & \vdots \\ \alpha_1\alpha_p & \alpha_2\alpha_p & \cdots & 1+\alpha_p^2 \end{pmatrix}$$

Now consider the product of a matrix and its transpose, B^TB below.

$$\begin{pmatrix} 1 & 0 & \cdots & 0 & \alpha_1 \\ 0 & 1 & & 0 & \alpha_2 \\ \vdots & & \ddots & & \vdots \\ 0 & & & 1 & \alpha_p \\ -\alpha_1 & -\alpha_2 & \cdots & -\alpha_p & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & \cdots & 0 & -\alpha_1 \\ 0 & 1 & & 0 & -\alpha_2 \\ \vdots & & \ddots & & \vdots \\ 0 & & & 1 & -\alpha_p \\ \alpha_1 & \alpha_2 & \cdots & \alpha_p & 1 \end{pmatrix} \tag{14.7}$$

This product equals a matrix of the form

$$\begin{pmatrix} A(\alpha_1,\cdots,\alpha_p) & \mathbf{0} \\ \mathbf{0} & 1+\sum_{i=1}^p \alpha_i^2 \end{pmatrix}$$

Therefore, $\left(1+\sum_{i=1}^p \alpha_i^2\right)\det(A(\alpha_1,\cdots,\alpha_p)) = \det(B)^2 = \det\left(B^T\right)^2$. However, using row operations,

$$\det B^T = \det\begin{pmatrix} 1 & 0 & \cdots & 0 & \alpha_1 \\ 0 & 1 & & 0 & \alpha_2 \\ \vdots & & \ddots & & \vdots \\ 0 & & & 1 & \alpha_p \\ 0 & 0 & \cdots & 0 & 1+\sum_{i=1}^p \alpha_i^2 \end{pmatrix} = 1+\sum_{i=1}^p \alpha_i^2$$

and therefore,

$$\left(1+\sum_{i=1}^p \alpha_i^2\right)\det(A(\alpha_1,\cdots,\alpha_p)) = \left(1+\sum_{i=1}^p \alpha_i^2\right)^2$$

which shows $\det(A(\alpha_1,\cdots,\alpha_p)) = \left(1+\sum_{i=1}^p \alpha_i^2\right)$. ■

Now consider the case of σ on ∂U. The maps will be of the form

$$\hat{\boldsymbol{x}} \in Q_k \to \begin{pmatrix} x_1 & \cdots & x_{i-1} & g(\hat{\boldsymbol{x}}_i) & x_{i+1} & \cdots & x_p \end{pmatrix}^T = \boldsymbol{h}(\hat{\boldsymbol{x}}_i)$$

I need to describe $\det\left(D\boldsymbol{h}(\hat{\boldsymbol{x}}_i)^* D\boldsymbol{h}(\hat{\boldsymbol{x}}_i)\right)^{1/2} \equiv J(\hat{\boldsymbol{x}})$.

Consider an example sufficient to see what happens in general in which $p=3$ and $i=2$. Then in this case, $J(\hat{\boldsymbol{x}})$ will be the square root of the determinant of

$$\begin{pmatrix} 1 & g_{x_1} & 0 \\ 0 & g_{x_3} & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ g_{x_1} & g_{x_3} \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} g_{x_1}^2+1 & g_{x_1}g_{x_3} \\ g_{x_1}g_{x_3} & g_{x_3}^2+1 \end{pmatrix}.$$

One can verify that this is just a special case in which $D\boldsymbol{h}(\hat{\boldsymbol{x}}_i)^* D\boldsymbol{h}(\hat{\boldsymbol{x}}_i)$ will be of the form considered in Lemma 14.4.2. Thus by this lemma, $J(\boldsymbol{x}) = \sqrt{1+\sum_{k\neq i} g_{,x_k}^2}$.

Then if $U\cap Q$ is of the form in 14.5 or in 14.6 one can identify the unit normal to the surface either on the top or the bottom of $U\cap Q$ from beginning calculus. These are respectively

$$\boldsymbol{n} = \frac{\begin{pmatrix} -g_{,x_1} & \cdots & -g_{,x_{p-1}} & \cdots & 1 \end{pmatrix}^T}{\sqrt{1+\sum_{k=1}^{p-1} g_{,x_k}^2}}, \frac{\begin{pmatrix} g_{,x_1} & \cdots & g_{,x_{p-1}} & \cdots & -1 \end{pmatrix}^T}{\sqrt{1+\sum_{k=1}^{p-1} g_{,x_k}^2}}$$

The first pointing up away from U and the second pointing down away from U.

If you simply assume g_k is differentiable, there is no problem in Definition 14.4.1. However, I will show later with Rademacher's theorem that it suffices to assume these functions are Lipschitz continuous.

In the following proof, I will regard $f(x_1,x_2,...,x_p)$ as a function of the listed variables.

Theorem 14.4.3 *Let U be a bounded open set in $\mathbb{R}^p$ satisfying the conditions of Definition 14.4.1 and let $\boldsymbol{F} \in C^1(\overline{U})$. Then*

$$\int_U \sum_{i=1}^p F_{i,i}(\boldsymbol{x})\, dm_p = \int_{\partial U} \boldsymbol{F}\cdot \boldsymbol{n} d\sigma_{p-1}$$

where $\boldsymbol{n}$ is the unit exterior normal to U.

Proof: Let $\mathrm{spt}(\psi_i)$ be a compact subset of Q_i and $\sum_{i=0}^N \psi_i = 1$ on $\bar{U}$ and each ψ_i is infinitely differentiable. This partition of unity exists by Lemma 12.6.4. There is an explicit description of the unit outer normal for each point of the boundary of U described in the proof of Theorem 14.4.3 in either of the two cases described in Definition 14.4.1 and illustrated in the above picture. Then

$$\begin{aligned}\int_U \sum_i F_{i,i}(\boldsymbol{x})\, dm_p &= \int_U \sum_i \sum_{k=0}^N (\psi_k F)_{i,i}(\boldsymbol{x})\, dm_p = \sum_i \sum_{k=0}^N \int_U (\psi_k F)_{i,i}(\boldsymbol{x})\, dm_p \\ &= \sum_{k=0}^N \int_{Q_k} \sum_i (\psi_k F)_{i,i}(\boldsymbol{x})\, dm_p \end{aligned} \tag{14.8}$$

Now consider one of the terms in the above. For the sake of simplicity assume $k = p$ so that the special direction corresponds to x_p. Also, I will assume that the function $g(\hat{\boldsymbol{x}})$ is on the top, so it is like the left picture in the above picture. A similar argument works if $g(\hat{\boldsymbol{x}})$ were on the bottom. Either way we can specify a unit exterior normal a.e. I will omit the subscript on g_k, Q_k, and ψ_k.

Case that $i < p$: Pick $i < p$. Letting $\hat{Q}$ be $(x_1,...,x_{p-1})$ where $\boldsymbol{x} \in Q$, For any i,

$$\int_Q (\psi_k F)_{i,i}\, dm_p = \int_{\hat{Q}} \int_{-\infty}^{g(x_1,...,x_{p-1})} (\psi_k F_i)_i\, dx_p d\hat{x} = \int_{\hat{Q}} \int_{-\infty}^0 D_i(\psi F_i)(\hat{\boldsymbol{x}}, y+g(\hat{\boldsymbol{x}}))\, dy d\hat{x} \tag{14.9}$$

Now for $i < p$, that in the integrand is not $\frac{\partial}{\partial x_i}(\psi F_i)(\hat{\boldsymbol{x}}, y+g(\hat{\boldsymbol{x}}))$. Indeed, by the chain rule,

$$\frac{\partial}{\partial x_i}(\psi F_i)(\hat{\boldsymbol{x}}, y+g(\hat{\boldsymbol{x}})) = D_i(\psi F_i)(\hat{\boldsymbol{x}}, y+g(\hat{\boldsymbol{x}})) + D_p(\psi F_i)(\hat{\boldsymbol{x}}, y+g(\hat{\boldsymbol{x}}))\frac{\partial g(\hat{\boldsymbol{x}})}{\partial x_i}$$

Since $\mathrm{spt}(\psi) \subseteq Q$, it follows that 14.9 reduces to

$$\int_{-\infty}^0 \int_{\hat{Q}} \frac{\partial}{\partial x_i}(\psi F_i)(\hat{\boldsymbol{x}}, y+g(\hat{\boldsymbol{x}}))\, d\hat{x} dy - \int_{\hat{Q}} \int_{-\infty}^0 D_p(\psi F_i)(\hat{\boldsymbol{x}}, y+g(\hat{\boldsymbol{x}}))\frac{\partial g(\hat{\boldsymbol{x}})}{\partial x_i} dy d\hat{x}$$

$$= 0 - \int_{\hat{Q}} (\psi F_i)(\hat{\boldsymbol{x}}, g(\hat{\boldsymbol{x}}))\, d\hat{x}$$

Case that $i = p$: In this case, 14.9 becomes $\int_{\hat{Q}} (\psi F_p)(\hat{x}, g(\hat{x}))\, d\hat{x}$. Recall how it was just shown that the unit normal is $\frac{\left(-g_{x_1}, \ldots, -g_{x_{p-1}}, 1\right)}{\sqrt{\sum_{i=1}^{p-1} g_{x_k}^2 + 1}}$ and $d\sigma = \sqrt{\sum_{i=1}^{p-1} g_{x_k}^2 + 1}\, dm_{p-1}$. Then the above reduces to $\int_{\partial(Q \cap U)} (\psi \boldsymbol{F}) \cdot \boldsymbol{n} d\sigma$. The same result will hold for all the Q_i. The sign changes if in the situation of 14.6. As to Q_0, $\int_{Q_0} \sum_i (\psi_0 F)_{i,i}(\boldsymbol{x})\, dm_p = 0$ because $\operatorname{spt}(\psi_0) \subseteq Q_0$. Returning to 14.8, it follows that

$$\int_U \sum_i F_{i,i}(\boldsymbol{x})\, dm_p = \sum_{k=0}^{N} \int_{Q_k} \sum_i (\psi_k F)_{i,i}(\boldsymbol{x})\, dm_p = \sum_{k=0}^{N} \int_{Q_k} \sum_i (\psi_k F)_{i,i}(\boldsymbol{x})\, dm_p$$

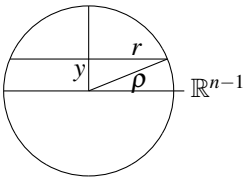
$$= \sum_{k=1}^{N} \int_{\partial(Q_k \cap U)} (\psi_k \boldsymbol{F}) \cdot \boldsymbol{n} d\sigma = \sum_{k=1}^{N} \int_{\partial U} (\psi_k \boldsymbol{F}) \cdot \boldsymbol{n} d\sigma$$

$$= \int_{\partial U} \left(\sum_{k=0}^{N} \psi_k \right) \boldsymbol{F} \cdot \boldsymbol{n} d\sigma = \int_{\partial U} \boldsymbol{F} \cdot \boldsymbol{n} d\sigma \ \blacksquare$$

14.5 Volumes of Balls in $\mathbb{R}^p$

Recall, $B(\boldsymbol{x}, r)$ denotes the set of all $\boldsymbol{y} \in \mathbb{R}^p$ such that $|\boldsymbol{y} - \boldsymbol{x}| < r$. By the change of variables formula for multiple integrals or simple geometric reasoning, all balls of radius r have the same volume. Furthermore, simple reasoning or change of variables formula will show that the volume of the ball of radius r equals $\alpha_p r^p$ where α_p will denote the volume of the unit ball in $\mathbb{R}^p$. With the divergence theorem, it is now easy to give a simple relationship between the surface area of the ball of radius r and the volume. Let dA denote the area measure called $d\alpha_{p-1}$ above. By the divergence theorem, $\int_{B(\mathbf{0},r)} \operatorname{div}(\boldsymbol{x})\, dx = \int_{\partial B(\mathbf{0},r)} \boldsymbol{x} \cdot \frac{\boldsymbol{x}}{|\boldsymbol{x}|} dA$ because the unit outward normal on $\partial B(\mathbf{0}, r)$ is $\frac{\boldsymbol{x}}{|\boldsymbol{x}|}$. Therefore, denoting $A(\partial B)$ as the area of ∂B, $p\alpha_p r^p = rA(\partial B(\mathbf{0}, r))$ and so $A(\partial B(\mathbf{0}, r)) = p\alpha_p r^{p-1}$. You recall the surface area of $S^2 \equiv \left\{ \boldsymbol{x} \in \mathbb{R}^3 : |\boldsymbol{x}| = r \right\}$ is given by $4\pi r^2$ while the volume of the ball, $B(\mathbf{0}, r)$ is $\frac{4}{3}\pi r^3$. This follows the above pattern. You just take the derivative with respect to the radius of the volume of the ball of radius r to get the area of the surface of this ball. Let ω_p denote the area of the sphere $S^{p-1} = \{\boldsymbol{x} \in \mathbb{R}^p : |\boldsymbol{x}| = 1\}$. I just showed that $\omega_p = p\alpha_p$.

I want to find α_p now.



Taking slices at height y as shown and using that these slices have $p-1$ dimensional area equal to $\alpha_{p-1} r^{p-1}$, it follows $\alpha_p \rho^p = 2\int_0^\rho \alpha_{p-1} \left(\rho^2 - y^2\right)^{(p-1)/2} dy$ since the r at a given y is $\sqrt{\rho^2 - y^2}$. In the integral, change variables, letting $y = \rho \cos\theta$. Then $\alpha_p \rho^p = 2\rho^p \alpha_{p-1} \int_0^{\pi/2} \sin^p(\theta)\, d\theta$. It follows that

$$\alpha_p = 2\alpha_{p-1} \int_0^{\pi/2} \sin^p(\theta)\, d\theta. \tag{14.10}$$

From this we find a formula for α_p.

First note that $\Gamma\left(\frac{1}{2}\right) = \int_0^\infty e^{-t}t^{-1/2}dt = \int_0^\infty e^{-u^2}u^{-1}2udu = 2\int_0^\infty e^{-u^2} = \sqrt{\pi}$ from elementary calculus using polar coordinates and change of variables. See Problem 2 on Page 361. Also see the problem on the gamma function $\Gamma(\alpha)\,\alpha = \Gamma(\alpha+1)$.

Theorem 14.5.1 *$\alpha_p = \frac{\pi^{p/2}}{\Gamma\left(\frac{p}{2}+1\right)}$ where Γ denotes the gamma function, defined for $\alpha > 0$ by $\Gamma(\alpha) \equiv \int_0^\infty e^{-t}t^{\alpha-1}dt$.*

Proof: Let $p = 1$ first. Then $\alpha_1 = \pi = \frac{\pi^{1/2}}{\Gamma\left(\frac{1}{2}+1\right)}$ because $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$ so the right side is $\frac{\pi^{1/2}}{\frac{1}{2}\Gamma\left(\frac{1}{2}\right)} = 2$ which is indeed the one dimensional area of the unit ball in one dimension. Similarly it is true for $p = 2,3$. Assume true for $p \geq 3$. Then using 14.10 and induction,

$$\alpha_{p+1} = 2\frac{\pi^{p/2}}{\Gamma\left(\frac{p}{2}+1\right)}\int_0^{\pi/2}\sin^{p+1}(\theta)\,d\theta$$

Using an integration by parts, this equals $2\frac{\pi^{p/2}}{\Gamma\left(\frac{p}{2}+1\right)}\frac{p}{p+1}\int_0^{\pi/2}\sin^{p-1}(\theta)\,d\theta$. By 14.10 and induction this is

$$2\frac{\pi^{p/2}}{\Gamma\left(\frac{p}{2}+1\right)}\frac{p}{p+1}\frac{\alpha_{p-1}}{\alpha_{p-2}} = 2\frac{\pi^{p/2}}{\Gamma\left(\frac{p}{2}+1\right)}\frac{p}{p+1}\frac{\frac{\pi^{(p-1)/2}}{\Gamma\left(\frac{p-1}{2}+1\right)}}{\frac{\pi^{(p-2)/2}}{\Gamma\left(\frac{p-2}{2}+1\right)}} =$$

$$2\frac{\pi^{(p+1)/2}}{\Gamma\left(\frac{p}{2}+1\right)}\frac{1}{p+1}\frac{p\Gamma\left(\frac{p}{2}\right)}{\Gamma\left(\frac{p+1}{2}\right)} = \frac{\pi^{(p+1)/2}}{\Gamma\left(\frac{p}{2}+1\right)}\frac{p\Gamma\left(\frac{p}{2}\right)}{\Gamma\left(\frac{p+1}{2}\right)\frac{p+1}{2}} = \frac{\pi^{(p+1)/2}}{\Gamma\left(\frac{p+1}{2}+1\right)}\blacksquare$$

14.6 Exercises

1. A random vector $\boldsymbol{X}$, with values in $\mathbb{R}^p$ has a multivariate normal distribution written as $\boldsymbol{X} \sim N_p(\boldsymbol{m},\Sigma)$ if for all Borel $E \subseteq \mathbb{R}^p$,

$$\lambda_{\boldsymbol{X}}(E) = \int_{\mathbb{R}^p}\mathcal{X}_E(\boldsymbol{x})\frac{1}{(2\pi)^{p/2}\det(\Sigma)^{1/2}}e^{\frac{-1}{2}(\boldsymbol{x}-\boldsymbol{m})^*\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{m})}dm_p$$

Here Σ is a positive definite symmetric matrix. Recall that

$$\lambda_{\boldsymbol{X}}(E) \equiv P(\boldsymbol{X} \in E).$$

Using the change of variables formula, show that $\lambda_{\boldsymbol{X}}$ defined above is a probability measure. One thing you must show is that

$$\int_{\mathbb{R}^p}\frac{1}{(2\pi)^{p/2}\det(\Sigma)^{1/2}}e^{\frac{-1}{2}(\boldsymbol{x}-\boldsymbol{m})^*\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{m})}dm_p = 1$$

Hint: To do this, you might use the fact from linear algebra that $\Sigma = Q^*DQ$ where D is a diagonal matrix and Q is an orthogonal matrix. Thus $\Sigma^{-1} = Q^*D^{-1}Q$. Maybe you could first let $\boldsymbol{y} = D^{-1/2}Q(\boldsymbol{x}-\boldsymbol{m})$ and change the variables. Note that the change of variables formula works fine when the open sets are all of $\mathbb{R}^p$. You don't need to confine your attention to finite open sets which would be the case with Riemann integrals which are only defined on bounded sets.

2. Consider the surface $z = x^2$ for $(x,y) \in (0,1) \times (0,1)$. Find the area of this surface. **Hint:** You can make do with just one chart in this case. Let $\boldsymbol{R}^{-1}(x,y) = \left(x,y,x^2\right)^T, (x,y) \in (0,1) \times (0,1)$. Then

$$D\boldsymbol{R}^{-1} = \begin{pmatrix} 1 & 0 & 2x \\ 0 & 1 & 0 \end{pmatrix}^T$$

It follows that $D\boldsymbol{R}^{-1*}D\boldsymbol{R}^{-1} = \begin{pmatrix} 4x^2+1 & 0 \\ 0 & 1 \end{pmatrix}$.

3. A parametrization for most of the sphere of radius $a > 0$ in three dimensions is

$$\begin{gathered} x = a\sin(\phi)\cos(\theta) \\ y = a\sin(\phi)\sin(\theta) \\ z = a\cos(\phi) \end{gathered}$$

where we will let $\phi \in (0,\pi), \theta \in (0,2\pi)$ so there is just one chart involved. As mentioned earlier, this includes all of the sphere except for the line of longitude corresponding to $\theta = 0$. Find a formula for the area of this sphere. Again, we are making do with a single chart.

4. For Ω a piecewise differentiable manifold and $f \geq 0$ and measurable, define $\int f d\sigma_p$ in terms of coordinate maps.

5. A topological manifold like that above will be called a differentiable manifold if there exists an atlas $\{(U_i, \boldsymbol{R}_i)\}$ for which the overlap maps $\boldsymbol{R}_i \circ \boldsymbol{R}_j^{-1}$ are differentiable. Then another atlas $\left\{(V_j, \boldsymbol{S}_j)\right\}$ will be equivalent to the given atlas if $\boldsymbol{S}_i \circ \boldsymbol{R}_j^{-1} : \boldsymbol{R}_j(V_i \cap U_j) \to \boldsymbol{S}_i(V_i \cap U_j)$ is differentiable. Then the given metric space with the given equivalence class of equivalent atlass will be called a differentiable manifold. Note how this approach avoids mentioning $\mathbb{R}^q$ as an underlying metric space for Ω. However, it is convenient to do it like I did in order to define $J_i(\boldsymbol{u})$ as above and tie it in to known material from calculus. Could you generalize having $\mathbb{R}^q$ contain Ω and still keep something like $J_i(\boldsymbol{u})$ which would give a formula for the surface measure on Ω? After all, the necessary machinery involving partitions of unity is available for metric space.

6. Let V be such that the divergence theorem holds. Show that $\int_V \nabla \cdot (u\nabla v)\, dV = \int_{\partial V} u \frac{\partial v}{\partial n}\, dA$ where $\boldsymbol{n}$ is the exterior normal and $\frac{\partial v}{\partial n}$ denotes the directional derivative of v in the direction $\boldsymbol{n}$**.** Remember the directional derivative.

$$\lim_{t\to 0} \frac{v(\boldsymbol{x}+t\boldsymbol{n}) - v(\boldsymbol{x})}{t} = \lim_{t\to 0} \frac{Dv(\boldsymbol{x})(t\boldsymbol{n}) + o(t)}{t} = Dv(\boldsymbol{x})(\boldsymbol{n}) = \nabla v(\boldsymbol{x}) \cdot \boldsymbol{n}$$

7. To prove the divergence theorem, it was shown first that the spacial partial derivative in the volume integral could be exchanged for multiplication by an appropriate component of the exterior normal. This problem starts with the divergence theorem and goes the other direction. Assuming the divergence theorem, holds for a region V, show that $\int_{\partial V} \boldsymbol{n} u\, dA = \int_V \nabla u\, dV$. Note this implies $\int_V \frac{\partial u}{\partial x}\, dV = \int_{\partial V} n_1 u\, dA$.

8. Fick's law for diffusion states the flux of a diffusing species, $\boldsymbol{J}$ is proportional to the gradient of the concentration c. Write this law getting the sign right for the constant of proportionality and derive an equation similar to the heat equation for the concentration c. Typically, c is the concentration of some sort of pollutant or a chemical.

9. Sometimes people consider diffusion in materials which are not homogeneous. This means that $\boldsymbol{J} = -K\nabla c$ where K is a 3×3 matrix and c is called the concentration. Thus in terms of components, $J_i = -\sum_j K_{ij}\frac{\partial c}{\partial x_j}$. Here c is the concentration which means the amount of pollutant or whatever is diffusing in a volume is obtained by integrating c over the volume. Derive a formula for a nonhomogeneous model of diffusion based on the above.

10. Let V be such that the divergence theorem holds. Show that

$$\int_V \left(v\nabla^2 u - u\nabla^2 v\right) dV = \int_{\partial V}\left(v\frac{\partial u}{\partial n} - u\frac{\partial v}{\partial n}\right) dA$$

where $\boldsymbol{n}$ is the exterior normal and $\frac{\partial u}{\partial n}$ is defined in Problem 6. Here $\nabla^2 u \equiv \sum_i u_{,x_i x_i}$.

11. Let V be a ball and suppose $\nabla^2 u = f$ in V while $u = g$ on ∂V. Show that there is at most one solution to this boundary value problem which is C^2 in V and continuous on V with its boundary. **Hint:** You might consider $w = u - v$ where u and v are solutions to the problem. Then use the result of Problem 6 and the identity $w\nabla^2 w = \nabla\cdot(w\nabla w) - \nabla w\cdot\nabla w$ to conclude $\nabla w = 0$. Then show this implies w must be a constant by considering $h(t) = w(t\boldsymbol{x} + (1-t)\boldsymbol{y})$ and showing h is a constant.

12. Show that $\int_{\partial V}\nabla\times\boldsymbol{v}\cdot\boldsymbol{n}\,dA = 0$ where V is a region for which the divergence theorem holds and $\boldsymbol{v}$ is a C^2 vector field.

13. Let $\boldsymbol{F}(x,y,z) = (x,y,z)$ be a vector field in $\mathbb{R}^3$ and let V be a three dimensional shape and let $\boldsymbol{n} = (n_1,n_2,n_3)$. Show that $\int_{\partial V}(xn_1 + yn_2 + zn_3)\,dA = 3\times$ volume of V.

14. Let $\boldsymbol{F} = x\boldsymbol{i} + y\boldsymbol{j} + z\boldsymbol{k}$ and let V denote the tetrahedron formed by the planes, $x = 0, y = 0, z = 0$, and $\frac{1}{3}x + \frac{1}{3}y + \frac{1}{5}z = 1$. Verify the divergence theorem for this example.

15. Suppose $f : U \to \mathbb{R}$ is continuous where U is some open set and for all $B \subseteq U$ where B is a ball, $\int_B f(\boldsymbol{x})\,dV = 0$. Show that this implies $f(\boldsymbol{x}) = 0$ for all $\boldsymbol{x}\in U$.

16. Let U denote the box centered at $(0,0,0)$ with sides parallel to the coordinate planes which has width 4, length 2 and height 3. Find the flux integral $\int_{\partial U}\boldsymbol{F}\cdot\boldsymbol{n}dS$ where $\boldsymbol{F} = (x+3, 2y, 3z)$. **Hint:** If you like, you might want to use the divergence theorem.

17. Find the flux out of the cylinder whose base is $x^2 + y^2 \le 1$ which has height 2 of the vector field $\boldsymbol{F} = \left(xy, zy, z^2 + x\right)$.

18. Find the flux out of the ball of radius 4 centered at $\boldsymbol{0}$ of the vector field $\boldsymbol{F} = (x, zy, z+x)$.

19. In one dimension, the heat equation is of the form $u_t = \alpha u_{xx}$. Show that $u(x,t) = e^{-\alpha n^2 t}\sin(nx)$ satisfies the heat equation

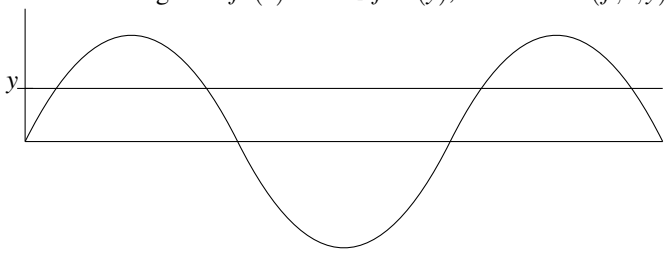
Chapter 15

Degree Theory

This chapter is on the Brouwer degree, a very useful concept with numerous and important applications. The degree can be used to prove some difficult theorems in topology such as the Brouwer fixed point theorem, the Jordan separation theorem, and the invariance of domain theorem. A couple of these big theorems have been presented earlier, but when you have degree theory, they get much easier. Degree theory is also used in bifurcation theory and many other areas in which it is an essential tool. The degree will be developed for $\mathbb{R}^p$ in this book. When this is understood, it is not too difficult to extend to versions of the degree which hold in Banach space. There is more on degree theory in the book by Deimling [12] and much of the presentation here follows this reference. Another more recent book which is really good is [15]. This is a whole book on degree theory.

The original reference for the approach given here, based on analysis, is [27] and dates from 1959. The degree was developed earlier by Brouwer and others using different methods. The more classical approach based on simplices and approximations with these is in [29]. I have given an approach based on singular homology as an appendix in [38].

To give you an idea what the degree is about, consider a real valued C^1 function defined on an interval I, and let $y \in f(I)$ be such that $f'(x) \neq 0$ for all $x \in f^{-1}(y)$. In this case the degree is the sum of the signs of $f'(x)$ for $x \in f^{-1}(y)$, written as $d(f, I, y)$.



In the above picture, $d(f, I, y)$ is 0 because there are two places where the sign is 1 and two where it is -1.

The amazing thing about this is the number you obtain in this simple manner is a specialization of something which is defined for continuous functions and which has nothing to do with differentiability. The reason one can extend the above simple idea to continuous functions is is an integral expression for the degree which is insensitive to homotopy. It is very similar to the winding number of complex analysis. The difference between the two is that with the degree, the integral which ties it all together is taken over the open set while the winding number is taken over the boundary, although proofs of it in the case of the winding number sometimes involve Green's theorem which involves an integral over the open set. I think these analogies are better seen in the other presentation in [38].

In this chapter Ω will refer to a bounded open set.

Definition 15.0.1 *For Ω a bounded open set, denote by $C(\overline{\Omega})$ the set of functions which are restrictions of functions in $C_c(\mathbb{R}^p)$, equivalently $C(\mathbb{R}^p)$ to $\overline{\Omega}$ and by $C^m(\overline{\Omega})$, $m \leq \infty$ the space of restrictions of functions in $C_c^m(\mathbb{R}^p)$, equivalently $C^m(\mathbb{R}^p)$ to $\overline{\Omega}$. If $f \in C(\overline{\Omega})$ the symbol f will also be used to denote a function defined on $\mathbb{R}^p$ equalling f on $\overline{\Omega}$ when convenient. The subscript c indicates that the functions have compact support. The norm in $C(\overline{\Omega})$ is defined as follows.*

$$\|f\|_{\infty,\overline{\Omega}} = \|f\|_\infty \equiv \sup\{|f(\boldsymbol{x})| : \boldsymbol{x} \in \overline{\Omega}\}.$$

If the functions take values in $\mathbb{R}^p$ write $C^m\left(\overline{\Omega};\mathbb{R}^p\right)$ or $C\left(\overline{\Omega};\mathbb{R}^p\right)$ for these functions if there is no differentiability assumed. The norm on $C\left(\overline{\Omega};\mathbb{R}^p\right)$ is defined in the same way as above,

$$\|\boldsymbol{f}\|_{\infty,\overline{\Omega}} = \|\boldsymbol{f}\|_{\infty} \equiv \sup\left\{|\boldsymbol{f}(\boldsymbol{x})| : \boldsymbol{x} \in \overline{\Omega}\right\}.$$

Of course if $m = \infty$, the notation means that there are infinitely many derivatives. Also, $C(\Omega;\mathbb{R}^p)$ consists of functions which are continuous on Ω that have values in $\mathbb{R}^p$ and $C^m(\Omega;\mathbb{R}^p)$ denotes the functions which have m continuous derivatives defined on Ω. Also let $\mathscr{P}$ consist of functions $\boldsymbol{f}(\boldsymbol{x})$ such that $f_k(\boldsymbol{x})$ is a polynomial, meaning an element of the algebra of functions generated by $\left\{1, x_1, \cdots, x_p\right\}$. Thus a typical polynomial is of the form $\sum_{i_1 \cdots i_p} a(i_1 \cdots i_p) x^{i_1} \cdots x^{i_p}$ where the i_j are nonnegative integers and $a(i_1 \cdots i_p)$ is a real number.

Some of the theorems are simpler if you base them on the Weierstrass approximation theorem.

Note that, by applying the Tietze extension theorem to the components of the function, one can always extend a function continuous on $\overline{\Omega}$ to all of $\mathbb{R}^p$ so there is no loss of generality in simply regarding functions continuous on $\overline{\Omega}$ as restrictions of functions continuous on $\mathbb{R}^p$. Next is the idea of a regular value.

Definition 15.0.2 *For W an open set in $\mathbb{R}^p$ and $\boldsymbol{g} \in C^1(W;\mathbb{R}^p)$ $\boldsymbol{y}$ is called a regular value of $\boldsymbol{g}$ if whenever $\boldsymbol{x} \in \boldsymbol{g}^{-1}(\boldsymbol{y})$, $\det(D\boldsymbol{g}(\boldsymbol{x})) \neq 0$. Note that if $\boldsymbol{g}^{-1}(\boldsymbol{y}) = \emptyset$, it follows that $\boldsymbol{y}$ is a regular value from this definition. That is, $\boldsymbol{y}$ is a regular value if and only if*

$$\boldsymbol{y} \notin \boldsymbol{g}(\{\boldsymbol{x} \in W : \det D\boldsymbol{g}(\boldsymbol{x}) = 0\})$$

Denote by $S_{\boldsymbol{g}}$ the set of singular values of $\boldsymbol{g}$, those $\boldsymbol{y}$ such that $\det(D\boldsymbol{g}(\boldsymbol{x})) = 0$ for some $\boldsymbol{x} \in \boldsymbol{g}^{-1}(\boldsymbol{y})$.

Also, $\partial\Omega$ will often be referred to. It is those points with the property that every open set (or open ball) containing the point contains points not in Ω and points in Ω. Then the following simple lemma will be used frequently.

Lemma 15.0.3 *Define ∂U to be those points $\boldsymbol{x}$ with the property that for every $r > 0$, $B(\boldsymbol{x},r)$ contains points of U and points of U^C. Then for U an open set, $\partial U = \overline{U} \setminus U$. Let C be a closed subset of $\mathbb{R}^p$ and let $\mathscr{K}$ denote the set of components of $\mathbb{R}^p \setminus C$. Then if K is one of these components, it is open and $\partial K \subseteq C$.*

Proof: Let $\boldsymbol{x} \in \overline{U} \setminus U$. If $B(\boldsymbol{x},r)$ contains no points of U, then $\boldsymbol{x} \notin \overline{U}$. If $B(\boldsymbol{x},r)$ contains no points of U^C, then $\boldsymbol{x} \in U$ and so $\boldsymbol{x} \notin \overline{U} \setminus U$. Therefore, $\overline{U} \setminus U \subseteq \partial U$. Now let $\boldsymbol{x} \in \partial U$. If $\boldsymbol{x} \in U$, then since U is open there is a ball containing $\boldsymbol{x}$ which is contained in U contrary to $\boldsymbol{x} \in \partial U$. Therefore, $\boldsymbol{x} \notin U$. If $\boldsymbol{x}$ is not a limit point of U, then some ball containing $\boldsymbol{x}$ contains no points of U contrary to $\boldsymbol{x} \in \partial U$. Therefore, $\boldsymbol{x} \in \overline{U} \setminus U$ which shows the two sets are equal.

Why is K open for K a component of $\mathbb{R}^p \setminus C$? This follows from Theorem 3.11.12 and results from open balls being connected. Thus if $k \in K$,letting $B(k,r) \subseteq C^C$, it follows $K \cup B(k,r)$ is connected and contained in C^C and therefore is contained in K because K is maximal with respect to being connected and contained in C^C.

Now for K a component of $\mathbb{R}^p \setminus C$, why is $\partial K \subseteq C$? Let $\boldsymbol{x} \in \partial K$. If $\boldsymbol{x} \notin C$, then $\boldsymbol{x} \in K_1$, some component of $\mathbb{R}^p \setminus C$. If $K_1 \neq K$ then $\boldsymbol{x}$ cannot be a limit point of K and so it cannot

be in ∂K. Therefore, $K = K_1$ but this also is a contradiction because if $\boldsymbol{x} \in \partial K$ then $\boldsymbol{x} \notin K$ thanks to the first part that $\partial U = \overline{U} \setminus U$. ■

Note that for an open set $U \subseteq \mathbb{R}^p$, and $\boldsymbol{h} : \overline{U} \to \mathbb{R}^p$, $\operatorname{dist}(\boldsymbol{h}(\partial U), \boldsymbol{y}) \geq \operatorname{dist}\left(\boldsymbol{h}\left(\overline{U}\right), \boldsymbol{y}\right)$ because $\overline{U} \supseteq \partial U$.

The following lemma will be nice to keep in mind.

Lemma 15.0.4 *$\boldsymbol{f} \in C\left(\overline{\Omega} \times [a,b]; \mathbb{R}^p\right)$ if and only if*

$$t \to \boldsymbol{f}(\cdot, t) \in C\left([a,b]; C\left(\overline{\Omega}; \mathbb{R}^p\right)\right)$$

Also

$$\|\boldsymbol{f}\|_{\infty, \overline{\Omega} \times [a,b]} = \max_{t \in [a,b]} \left(\|\boldsymbol{f}(\cdot, t)\|_{\infty, \overline{\Omega}} \right)$$

Proof: $\Rightarrow$By uniform continuity, if $\varepsilon > 0$ there is $\delta > 0$ such that if $|t - s| < \delta$, then for all $\boldsymbol{x} \in \overline{\Omega}$, $\|\boldsymbol{f}(\boldsymbol{x}, t) - \boldsymbol{f}(\boldsymbol{x}, s)\| < \frac{\varepsilon}{2}$. It follows that

$$\|\boldsymbol{f}(\cdot, t) - \boldsymbol{f}(\cdot, s)\|_{\infty} \leq \frac{\varepsilon}{2} < \varepsilon$$

$\Leftarrow$Say $(\boldsymbol{x}_n, t_n) \to (\boldsymbol{x}, t)$. Does it follow that $\boldsymbol{f}(\boldsymbol{x}_n, t_n) \to \boldsymbol{f}(\boldsymbol{x}, t)$?

$$\begin{aligned} \|\boldsymbol{f}(\boldsymbol{x}_n, t_n) - \boldsymbol{f}(\boldsymbol{x}, t)\| &\leq \|\boldsymbol{f}(\boldsymbol{x}_n, t_n) - \boldsymbol{f}(\boldsymbol{x}_n, t)\| + \|\boldsymbol{f}(\boldsymbol{x}_n, t) - \boldsymbol{f}(\boldsymbol{x}, t)\| \\ &\leq \|\boldsymbol{f}(\cdot, t_n) - \boldsymbol{f}(\cdot, t)\|_{\infty} + \|\boldsymbol{f}(\boldsymbol{x}_n, t) - \boldsymbol{f}(\boldsymbol{x}, t)\| \end{aligned}$$

both terms converge to 0, the first because $\boldsymbol{f}$ is continuous into $C\left(\overline{\Omega}; \mathbb{R}^p\right)$ and the second because $\boldsymbol{x} \to \boldsymbol{f}(\boldsymbol{x}, t)$ is continuous.

The claim about the norms is next. Let $(\boldsymbol{x}, t)$ be such that $\|\boldsymbol{f}\|_{\infty, \overline{\Omega} \times [a,b]} < \|\boldsymbol{f}(\boldsymbol{x}, t)\| + \varepsilon$. Then

$$\|\boldsymbol{f}\|_{\infty, \overline{\Omega} \times [a,b]} < \|\boldsymbol{f}(\boldsymbol{x}, t)\| + \varepsilon \leq \max_{t \in [a,b]} \left(\|\boldsymbol{f}(\cdot, t)\|_{\infty, \overline{\Omega}} \right) + \varepsilon$$

and so $\|\boldsymbol{f}\|_{\infty, \overline{\Omega} \times [a,b]} \leq \max_{t \in [a,b]} \max \left(\|\boldsymbol{f}(\cdot, t)\|_{\infty, \overline{\Omega}} \right)$ because ε is arbitrary. However, the same argument works in the other direction. There exists t such that

$$\|\boldsymbol{f}(\cdot, t)\|_{\infty, \overline{\Omega}} = \max_{t \in [a,b]} \left(\|\boldsymbol{f}(\cdot, t)\|_{\infty, \overline{\Omega}} \right)$$

by compactness of the interval. Then by compactness of $\overline{\Omega}$, there is $\boldsymbol{x}$ such that

$$\|\boldsymbol{f}(\cdot, t)\|_{\infty, \overline{\Omega}} = \|\boldsymbol{f}(\boldsymbol{x}, t)\| \leq \|\boldsymbol{f}\|_{\infty, \overline{\Omega} \times [a,b]}$$

and so the two norms are the same. ■

15.1 Sard's Lemma and Approximation

First are easy assertions about approximation of continuous functions with smooth ones.

The following is the Weierstrass approximation theorem. It is Corollary 5.5.3 presented earlier.

Corollary 15.1.1 *If $f \in C([a,b];X)$ where X is a normed linear space, then there exists a sequence of polynomials which converge uniformly to f on $[a,b]$. The polynomials are of the form*

$$\sum_{k=0}^{m} \binom{m}{k} \left(l^{-1}(t)\right)^k \left(1-l^{-1}(t)\right)^{m-k} f\left(l\left(\frac{k}{m}\right)\right) \tag{15.1}$$

where l is a linear one to one and onto map from $[0,1]$ to $[a,b]$.

Applying the Weierstrass approximation theorem, Theorem 5.7.7 or Theorem 5.9.5 to the components of a vector valued function yields the following Theorem.

Theorem 15.1.2 *If $\boldsymbol{f} \in C\left(\overline{\Omega};\mathbb{R}^p\right)$ for Ω a bounded subset of $\mathbb{R}^p$, then for all $\varepsilon > 0$, there exists $\boldsymbol{g} \in C^\infty\left(\overline{\Omega};\mathbb{R}^p\right)$ such that $\|\boldsymbol{g}-\boldsymbol{f}\|_{\infty,\overline{\Omega}} < \varepsilon$.*

Recall Sard's lemma, shown earlier. It is Lemma 11.14.4. I am stating it here for convenience.

Lemma 15.1.3 *(Sard) Let Ω be an open set in $\mathbb{R}^p$ and let $\boldsymbol{h}: \Omega \to \mathbb{R}^p$ be differentiable. Let*

$$S \equiv \{\boldsymbol{x} \in \Omega : \det D\boldsymbol{h}(\boldsymbol{x}) = 0\}.$$

Then $m_p(\boldsymbol{h}(S)) = 0$.

First note that if $\boldsymbol{y} \notin \boldsymbol{g}(\Omega)$, then $\boldsymbol{y} \notin \boldsymbol{g}(\{\boldsymbol{x} \in \Omega : \det D\boldsymbol{g}(\boldsymbol{x}) = 0\})$ so it is a regular value.

Observe that any uncountable set in $\mathbb{R}^p$ has a limit point. To see this, tile $\mathbb{R}^p$ with countably many congruent boxes. One of them has uncountably many points. Now subdivide this into 2^p congruent boxes. One has uncountably many points. Continue subdividing this way to obtain a limit point as the unique point in the intersection of a nested sequence of compact sets whose diameters converge to 0.

Lemma 15.1.4 *Let $\boldsymbol{g} \in C^\infty(\mathbb{R}^p;\mathbb{R}^p)$ and let $\{\boldsymbol{y}_i\}_{i=1}^\infty$ be points of $\mathbb{R}^p$ and let $\eta > 0$. Then there exists $\boldsymbol{e}$ with $\|\boldsymbol{e}\| < \eta$ and $\boldsymbol{y}_i + \boldsymbol{e}$ is a regular value for $\boldsymbol{g}$ for all i.*

Proof: Let $S = \{\boldsymbol{x} \in \mathbb{R}^p : \det D\boldsymbol{g}(\boldsymbol{x}) = 0\}$. By Sard's lemma, $\boldsymbol{g}(S)$ has measure zero. Let $N \equiv \cup_{i=1}^\infty (\boldsymbol{g}(S) - \boldsymbol{y}_i)$. Thus N has measure 0. Pick $\boldsymbol{e} \in B(\boldsymbol{0},\eta) \setminus N$. Then for each $i, \boldsymbol{y}_i + \boldsymbol{e} \notin \boldsymbol{g}(S)$. ■

Next we approximate $\boldsymbol{f}$ with a smooth function $\boldsymbol{g}$ such that each $\boldsymbol{y}_i$ is a regular value of $\boldsymbol{g}$.

Lemma 15.1.5 *Let $\boldsymbol{f} \in C\left(\overline{\Omega};\mathbb{R}^p\right)$ and let $\{\boldsymbol{y}_i\}_{i=1}^\infty$ be points not in $\boldsymbol{f}(\partial\Omega)$ and let $\delta > 0$. Then there exists $\boldsymbol{g} \in C^\infty\left(\overline{\Omega};\mathbb{R}^p\right)$ such that $\|\boldsymbol{g}-\boldsymbol{f}\|_{\infty,\overline{\Omega}} < \delta$ and $\boldsymbol{y}_i$ is a regular value for $\boldsymbol{g}$ for each i. That is, **if** $\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{y}_i$, then $D\boldsymbol{g}(\boldsymbol{x})^{-1}$ exists. Also, if $\delta < \operatorname{dist}(\boldsymbol{f}(\partial\Omega),\boldsymbol{y})$ for some $\boldsymbol{y}$ a regular value of $\boldsymbol{g} \in C^\infty\left(\overline{\Omega};\mathbb{R}^p\right)$, then $\boldsymbol{g}^{-1}(\boldsymbol{y})$ is a finite set of points in Ω. Also, if $\boldsymbol{y}$ is a regular value of $\boldsymbol{g} \in C^\infty(\mathbb{R}^p,\mathbb{R}^p)$, then $\boldsymbol{g}^{-1}(\boldsymbol{y})$ is countable.*

Proof: Pick $\tilde{\boldsymbol{g}} \in C^\infty\left(\overline{\Omega};\mathbb{R}^p\right), \|\tilde{\boldsymbol{g}}-\boldsymbol{f}\|_{\infty,\overline{\Omega}} < \delta$. $\boldsymbol{g} \equiv \tilde{\boldsymbol{g}} - \boldsymbol{e}$ where $\boldsymbol{e}$ is from the above Lemma 15.1.4, $\|\boldsymbol{e}\| < \eta$ and η so small that $\|\boldsymbol{g}-\boldsymbol{f}\|_{\infty,\overline{\Omega}} < \delta$. Then if $\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{y}_i$, you get $\tilde{\boldsymbol{g}}(\boldsymbol{x}) = \boldsymbol{y}_i + \boldsymbol{e}$ with $\boldsymbol{x}$ a regular value of $\tilde{\boldsymbol{g}}$ and so $\det(D\boldsymbol{g}(\boldsymbol{x})) = \det(D\tilde{\boldsymbol{g}}(\boldsymbol{x})) \neq 0$ so this shows the first part.

It remains to verify the last claims. Since $\|g-f\|_{\overline{\Omega},\infty} < \delta$, if $x \in \partial\Omega$, then

$$\|g(x)-y\| \geq \|f(x)-y\| - \|f(x)-g(x)\| \geq \operatorname{dist}(f(\partial\Omega),y) - \delta > \delta - \delta = 0$$

and so $y \notin g(\partial\Omega)$, so if $g(x) = y$, then $x \in \Omega$. If there are infinitely many points in $g^{-1}(y)$, then there would be a subsequence converging to a point $x \in \overline{\Omega}$. Thus $g(x) = y$ so $x \notin \partial\Omega$. However, this would violate the inverse function theorem because g would fail to be one to one on an open ball containing x and contained in Ω. Therefore, there are only finitely many points in $g^{-1}(y)$ and at each point, the determinant of the derivative of g is nonzero. For y a regular value, $g^{-1}(y)$ is countable since otherwise, there would be a limit point $x \in g^{-1}(y)$ and g would fail to be one to one near x contradicting the inverse function theorem. ■

Now with this, here is a definition of the degree.

Definition 15.1.6 *Let Ω be a bounded open set in $\mathbb{R}^p$ and let $f : \overline{\Omega} \to \mathbb{R}^p$ be continuous. Let $y \notin f(\partial\Omega)$. Then the degree is defined as follows: Let g be infinitely differentiable,*

$$\|f-g\|_{\infty,\overline{\Omega}} < \delta \equiv \operatorname{dist}(f(\partial\Omega),y),$$

and y is a regular value of g.Then $y \notin g(\partial\Omega)$ and we define

$$d(f,\Omega,y) \equiv \sum \left\{ \operatorname{sgn}(\det(Dg(x))) : x \in g^{-1}(y), x \in \Omega \right\}$$

where the sum is finite by Lemma 15.1.5, defined to equal 0 if $g^{-1}(y)$ is empty.

Note that if g is such an approximation of f then if $x \in \partial\Omega$ and $t \in [0,1]$,

$$\begin{aligned} |tg(x) + (1-t)f(x) - y| &\geq |f(x)-y| - t\|g-f\|_\infty \\ &> \operatorname{dist}(f(\partial\Omega),y) - \operatorname{dist}(f(\partial\Omega),y) = 0 \end{aligned}$$

Thus $tg + (1-t)f$ maps no point of $\partial\Omega$ to y. In particular, g maps no point of $\partial\Omega$ to y.

Lemma 15.1.7 *The above definition makes sense for a single g and, assuming this is well defined, then it would follow that if $y \notin f(\Omega)$, then $d(f,\Omega,y) = 0$.*

Proof: As just noted if $\|f-g\|_{\infty,\overline{\Omega}} < \operatorname{dist}(f(\partial\Omega),y)$ then $y \notin g(\partial\Omega)$. In fact

$$y \notin (tg(x) + (1-t)f(x))(\partial\Omega)$$

for any $t \in [0,1]$. Thus the sum is a finite sum thanks to the inverse function theorem and the fact that each point in $g^{-1}(y)$ is a regular point. What if $y \notin f(\Omega)$? In this case, assuming the definition is well defined, you could pick g such that y is a regular value for g and also $\|f-g\|_{\infty,\overline{\Omega}} < \operatorname{dist}\left(f\left(\overline{\Omega}\right),y\right)$

$$\begin{aligned} \|g(x)-y\| &= \|(y-f(x)) - (g(x)-f(x))\| \geq \|f(x)-y\| - \|g(x)-f(x)\| \\ &> \operatorname{dist}\left(y,f\left(\overline{\Omega}\right)\right) - \operatorname{dist}\left(y,f\left(\overline{\Omega}\right)\right) = 0 \end{aligned}$$

so the above definition would say that $d(f,\Omega,y) = 0$ because there would be no terms in the sum. ■

We really need to verify that this definition is well defined. This involves the use of an integral.

Next is an identity. It was Lemma 7.11.2 on Page 204.

Lemma 15.1.8 *Let $\boldsymbol{g}:\Omega\to\mathbb{R}^p$ be C^2 where Ω is an open subset of $\mathbb{R}^p$. Then*

$$\sum_{j=1}^{p}\operatorname{cof}(D\boldsymbol{g})_{ij,j}=0,$$

where here $(D\boldsymbol{g})_{ij}\equiv g_{i,j}\equiv\frac{\partial g_i}{\partial x_j}$. Also, $\operatorname{cof}(D\boldsymbol{g})_{ij}=\frac{\partial\det(D\boldsymbol{g})}{\partial g_{i,j}}$.

Next is an integral representation of $\sum\left\{\operatorname{sgn}(\det(D\boldsymbol{g}(\boldsymbol{x}))):\boldsymbol{x}\in\boldsymbol{g}^{-1}(\boldsymbol{y})\right\}$ but first is a little lemma about disjoint sets.

Lemma 15.1.9 *Let K be a compact set and C a closed set in $\mathbb{R}^p$ such that $K\cap C=\emptyset$. Then*

$$\operatorname{dist}(K,C)\equiv\inf\{\|\boldsymbol{k}-\boldsymbol{c}\|:\boldsymbol{k}\in K,\boldsymbol{c}\in C\}>0.$$

Proof: Let $d\equiv\inf\{\|\boldsymbol{k}-\boldsymbol{c}\|:\boldsymbol{k}\in K,\boldsymbol{c}\in C\}$. Let $\{\boldsymbol{k}_i\},\{\boldsymbol{c}_i\}$ be such that

$$d+\frac{1}{i}>\|\boldsymbol{k}_i-\boldsymbol{c}_i\|.$$

Since K is compact, there is a subsequence still denoted by $\{\boldsymbol{k}_i\}$ such that $\boldsymbol{k}_i\to\boldsymbol{k}\in K$. Then also

$$\|\boldsymbol{c}_i-\boldsymbol{c}_m\|\le\|\boldsymbol{c}_i-\boldsymbol{k}_i\|+\|\boldsymbol{k}_i-\boldsymbol{k}_m\|+\|\boldsymbol{c}_m-\boldsymbol{k}_m\|$$

If $d=0$, then as $m,i\to\infty$ it follows $\|\boldsymbol{c}_i-\boldsymbol{c}_m\|\to0$ and so $\{\boldsymbol{c}_i\}$ is a Cauchy sequence which must converge to some $\boldsymbol{c}\in C$. But then $\|\boldsymbol{c}-\boldsymbol{k}\|=\lim_{i\to\infty}\|\boldsymbol{c}_i-\boldsymbol{k}_i\|=0$ and so $\boldsymbol{c}=\boldsymbol{k}\in C\cap K$, a contradiction to these sets being disjoint. ■

In particular the distance between a point and a closed set is always positive if the point is not in the closed set. Of course this is obvious even without the above lemma.

Definition 15.1.10 *Let $\boldsymbol{g}\in C^\infty(\overline{\Omega};\mathbb{R}^p)$ where Ω is a bounded open set. Also let ϕ_ε be a mollifier.*

$$\phi_\varepsilon\in C_c^\infty(B(\boldsymbol{0},\varepsilon)),\ \phi_\varepsilon\ge0,\ \int\phi_\varepsilon dx=1.$$

The idea is that ε will converge to 0 to get suitable approximations.

First, here is a technical lemma which will be used to identify the degree with an integral.

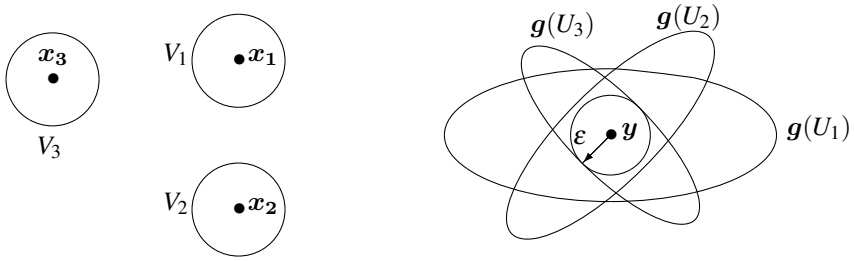
Lemma 15.1.11 *Let $\boldsymbol{y}\notin\boldsymbol{g}(\partial\Omega)$ for $\boldsymbol{g}\in C^\infty(\overline{\Omega};\mathbb{R}^p)$. Also suppose $\boldsymbol{y}$ is a regular value of $\boldsymbol{g}$. Then for all positive ε small enough,*

$$\int_\Omega\phi_\varepsilon(\boldsymbol{g}(\boldsymbol{x})-\boldsymbol{y})\det D\boldsymbol{g}(\boldsymbol{x})\,dx=\sum\left\{\operatorname{sgn}(\det D\boldsymbol{g}(\boldsymbol{x})):\boldsymbol{x}\in\boldsymbol{g}^{-1}(\boldsymbol{y})\right\}$$

Proof: First note that the sum is finite from Lemma 15.1.5. It only remains to verify the equation. If $\boldsymbol{y}\notin\boldsymbol{g}(\Omega)$, then for $\varepsilon<\operatorname{dist}(\boldsymbol{g}(\overline{\Omega}),\boldsymbol{y})$, $\phi_\varepsilon(\boldsymbol{g}(\boldsymbol{x})-\boldsymbol{y})=0$ for all $\boldsymbol{x}\in\Omega$ so both sides equal 0.

I need to show the left side of this equation is constant for ε small enough and equals the right side. By what was just shown, there are finitely many points, $\{\boldsymbol{x}_i\}_{i=1}^m=\boldsymbol{g}^{-1}(\boldsymbol{y})$. By the inverse function theorem, there exist disjoint open sets U_i with $\boldsymbol{x}_i\in U_i$, such that $\boldsymbol{g}$ is one

to one on U_i with $\det(D\boldsymbol{g}(\boldsymbol{x}))$ having constant sign on U_i and $\boldsymbol{g}(U_i)$ is an open set containing $\boldsymbol{y}$. Then let ε be small enough that $B(\boldsymbol{y},\varepsilon) \subseteq \cap_{i=1}^m \boldsymbol{g}(U_i)$. Also, $\boldsymbol{y} \notin \boldsymbol{g}\left(\overline{\Omega} \setminus \left(\cup_{i=1}^n U_i\right)\right)$, a compact set. Let ε be still smaller, if necessary, so that $B(\boldsymbol{y},\varepsilon) \cap \boldsymbol{g}\left(\overline{\Omega} \setminus \left(\cup_{i=1}^n U_i\right)\right) = \emptyset$ and let $V_i \equiv \boldsymbol{g}^{-1}(B(\boldsymbol{y},\varepsilon)) \cap U_i$.



Therefore, for any ε this small,

$$\int_\Omega \phi_\varepsilon(\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{y}) \det D\boldsymbol{g}(\boldsymbol{x})\, dx = \sum_{i=1}^m \int_{V_i} \phi_\varepsilon(\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{y}) \det D\boldsymbol{g}(\boldsymbol{x})\, dx$$

The reason for this is as follows. The integrand on the left is nonzero only if $\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{y} \in B(\boldsymbol{0},\varepsilon)$ which occurs only if $\boldsymbol{g}(\boldsymbol{x}) \in B(\boldsymbol{y},\varepsilon)$ which is the same as $\boldsymbol{x} \in \boldsymbol{g}^{-1}(B(\boldsymbol{y},\varepsilon))$. Therefore, the integrand is nonzero only if $\boldsymbol{x}$ is contained in exactly one of the disjoint sets, V_i. Now using the change of variables theorem, $(\boldsymbol{z} = \boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{y}, \boldsymbol{g}^{-1}(\boldsymbol{y}+\boldsymbol{z}) = \boldsymbol{x}.)$

$$= \sum_{i=1}^m \int_{\boldsymbol{g}(V_i)-\boldsymbol{y}} \phi_\varepsilon(\boldsymbol{z}) \det D\boldsymbol{g}\left(\boldsymbol{g}^{-1}(\boldsymbol{y}+\boldsymbol{z})\right) \left|\det D\boldsymbol{g}^{-1}(\boldsymbol{y}+\boldsymbol{z})\right| dz$$

By the chain rule, $I = D\boldsymbol{g}\left(\boldsymbol{g}^{-1}(\boldsymbol{y}+\boldsymbol{z})\right) D\boldsymbol{g}^{-1}(\boldsymbol{y}+\boldsymbol{z})$ and so

$$\begin{aligned}
&\det D\boldsymbol{g}\left(\boldsymbol{g}^{-1}(\boldsymbol{y}+\boldsymbol{z})\right) \left|\det D\boldsymbol{g}^{-1}(\boldsymbol{y}+\boldsymbol{z})\right| \\
&= \operatorname{sgn}\left(\det D\boldsymbol{g}\left(\boldsymbol{g}^{-1}(\boldsymbol{y}+\boldsymbol{z})\right)\right) \left|\det D\boldsymbol{g}\left(\boldsymbol{g}^{-1}(\boldsymbol{y}+\boldsymbol{z})\right)\right| \left|\det D\boldsymbol{g}^{-1}(\boldsymbol{y}+\boldsymbol{z})\right| \\
&= \operatorname{sgn}\left(\det D\boldsymbol{g}\left(\boldsymbol{g}^{-1}(\boldsymbol{y}+\boldsymbol{z})\right)\right) \\
&= \operatorname{sgn}(\det D\boldsymbol{g}(\boldsymbol{x})) = \operatorname{sgn}(\det D\boldsymbol{g}(\boldsymbol{x}_i)).
\end{aligned}$$

Therefore, this reduces to

$$\sum_{i=1}^m \operatorname{sgn}(\det D\boldsymbol{g}(\boldsymbol{x}_i)) \int_{\boldsymbol{g}(V_i)-\boldsymbol{y}} \phi_\varepsilon(\boldsymbol{z})\, dz =$$

$$\sum_{i=1}^m \operatorname{sgn}(\det D\boldsymbol{g}(\boldsymbol{x}_i)) \int_{B(\boldsymbol{0},\varepsilon)} \phi_\varepsilon(\boldsymbol{z})\, dz = \sum_{i=1}^m \operatorname{sgn}(\det D\boldsymbol{g}(\boldsymbol{x}_i)).$$

In case $\boldsymbol{g}^{-1}(\boldsymbol{y}) = \emptyset$, there exists $\varepsilon > 0$ such that $\boldsymbol{g}\left(\overline{\Omega}\right) \cap B(\boldsymbol{y},\varepsilon) = \emptyset$ and so for ε this small,

$$\int_\Omega \phi_\varepsilon(\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{y}) \det D\boldsymbol{g}(\boldsymbol{x})\, dx = 0. \blacksquare$$

As noted above, this will end up being $d(\boldsymbol{g},\Omega,\boldsymbol{y})$ in this last case where $\boldsymbol{g}^{-1}(\boldsymbol{y}) = \emptyset$.

Lemma 15.1.12 *Suppose $\boldsymbol{g},\hat{\boldsymbol{g}}$ both satisfy Definition 15.1.6. For δ given there, $\delta = \operatorname{dist}(\boldsymbol{f}(\partial\Omega),\boldsymbol{y})$,*

$$\delta > \|\boldsymbol{f}-\boldsymbol{g}\|_{\infty,\overline{\Omega}},\ \delta > \|\boldsymbol{f}-\hat{\boldsymbol{g}}\|_{\infty,\overline{\Omega}}$$

Then for $t\in[0,1]$ so does $t\boldsymbol{g}+(1-t)\hat{\boldsymbol{g}}$. In particular, $\boldsymbol{y}\notin(t\boldsymbol{g}+(1-t)\hat{\boldsymbol{g}})(\partial\Omega)$. Also $d(\boldsymbol{f}-\boldsymbol{y},\Omega,\boldsymbol{0}) = d(\boldsymbol{f},\Omega,\boldsymbol{y})$.

Proof: This follows from the fact that $B(\boldsymbol{y},\delta)$ in $\|\cdot\|_{\infty,\overline{\Omega}}$ is convex. From the triangle inequality, if $t\in[0,1]$,

$$\begin{aligned}\|\boldsymbol{f}-(t\boldsymbol{g}+(1-t)\hat{\boldsymbol{g}})\|_\infty &\leq t\|\boldsymbol{f}-\boldsymbol{g}\|_\infty+(1-t)\|\boldsymbol{f}-\hat{\boldsymbol{g}}\|_\infty\\ &< t\delta+(1-t)\delta=\delta.\end{aligned}$$

If $\|\boldsymbol{h}-\boldsymbol{f}\|_\infty<\delta$, as was just shown for $\boldsymbol{h}\equiv t\boldsymbol{g}+(1-t)\hat{\boldsymbol{g}}$, then if $\boldsymbol{x}\in\partial\Omega$,

$$\begin{aligned}\|\boldsymbol{y}-\boldsymbol{h}(\boldsymbol{x})\| &\geq \|\boldsymbol{y}-\boldsymbol{f}(\boldsymbol{x})\|-\|\boldsymbol{h}(\boldsymbol{x})-\boldsymbol{f}(\boldsymbol{x})\|\\ &> \operatorname{dist}(\boldsymbol{f}(\partial\Omega),\boldsymbol{y})-\delta\geq\delta-\delta=0\end{aligned}$$

Now consider the last claim. This follows because $\|\boldsymbol{g}-\boldsymbol{f}\|_\infty$ small is the same as $\|\boldsymbol{g}-\boldsymbol{y}-(\boldsymbol{f}-\boldsymbol{y})\|_\infty$ being small. They are the same. Also, $(\boldsymbol{g}-\boldsymbol{y})^{-1}(\boldsymbol{0}) = \boldsymbol{g}^{-1}(\boldsymbol{y})$ and $D\boldsymbol{g}(\boldsymbol{x}) = D(\boldsymbol{g}-\boldsymbol{y})(\boldsymbol{x})$. ■

Next is an important result on homotopy which is used to show that the above definition of the degree is well defined.

Lemma 15.1.13 *If $\boldsymbol{h}$ is in $C^\infty(\overline{\Omega}\times[a,b],\mathbb{R}^p)$, and $\boldsymbol{0}\notin\boldsymbol{h}(\partial\Omega\times[a,b])$ then for $0<\varepsilon<\operatorname{dist}(\boldsymbol{0},\boldsymbol{h}(\partial\Omega\times[a,b]))$,*

$$t\to\int_\Omega\phi_\varepsilon(\boldsymbol{h}(\boldsymbol{x},t))\det D_1\boldsymbol{h}(\boldsymbol{x},t)\,dx$$

is constant for $t\in[a,b]$. As a special case, $d(\boldsymbol{f},\Omega,\boldsymbol{y})$ is well defined. Also, if $\boldsymbol{y}\notin\boldsymbol{f}(\overline{\Omega})$, then $d(\boldsymbol{f},\Omega,\boldsymbol{y})=0$.

Proof: By continuity of $\boldsymbol{h}$, $\boldsymbol{h}(\partial\Omega\times[a,b])$ is compact and so is at a positive distance from $\boldsymbol{0}$. Let $\varepsilon>0$ be such that for all $t\in[a,b]$,

$$B(\boldsymbol{0},\varepsilon)\cap\boldsymbol{h}(\partial\Omega\times[a,b])=\emptyset \tag{15.2}$$

Define for $t\in(a,b)$,

$$H(t)\equiv\int_\Omega\phi_\varepsilon(\boldsymbol{h}(\boldsymbol{x},t))\det D_1\boldsymbol{h}(\boldsymbol{x},t)\,dx$$

I will show that $H'(t)=0$ on (a,b). Then, since H is continuous on $[a,b]$, it will follow from the mean value theorem that $H(t)$ is constant on $[a,b]$. If $t\in(a,b)$,

$$H'(t)=\int_\Omega\sum_\alpha\phi_{\varepsilon,\alpha}(\boldsymbol{h}(\boldsymbol{x},t))h_{\alpha,t}(\boldsymbol{x},t)\det D_1\boldsymbol{h}(\boldsymbol{x},t)\,dx$$

$$+\int_\Omega\phi_\varepsilon(\boldsymbol{h}(\boldsymbol{x},t))\sum_{\alpha,j}\det D_1(\boldsymbol{h}(\boldsymbol{x},t))_{,\alpha j}h_{\alpha,jt}dx\equiv\boldsymbol{A}+\boldsymbol{B}. \tag{15.3}$$

In this formula, the function det is considered as a function of the n^2 entries in the $n \times n$ matrix and the $,\alpha j$ represents the derivative with respect to the αj^{th} entry $h_{\alpha,j}$. Now as in the proof of Lemma 7.11.2 on Page 204,

$$\det D_1(\boldsymbol{h}(\boldsymbol{x},t))_{,\alpha j} = (\operatorname{cof} D_1(\boldsymbol{h}(\boldsymbol{x},t)))_{\alpha j}$$

and so

$$\boldsymbol{B} = \int_\Omega \sum_\alpha \sum_j \phi_\varepsilon(\boldsymbol{h}(\boldsymbol{x},t))(\operatorname{cof} D_1(\boldsymbol{h}(\boldsymbol{x},t)))_{\alpha j} h_{\alpha,jt} dx.$$

By hypothesis

$$\boldsymbol{x} \to \phi_\varepsilon(\boldsymbol{h}(\boldsymbol{x},t))(\operatorname{cof} D_1(\boldsymbol{h}(\boldsymbol{x},t)))_{\alpha j} \text{ for } \boldsymbol{x} \in \Omega$$

is in $C_c^\infty(\Omega)$ because if $\boldsymbol{x} \in \partial\Omega$, it follows that for all $t \in [a,b]$, $\boldsymbol{h}(\boldsymbol{x},t) \notin B(\boldsymbol{0},\varepsilon)$ and so $\phi_\varepsilon(\boldsymbol{h}(\boldsymbol{x},t)) = 0$ off some compact set contained in Ω. Therefore, integrate by parts and write

$$\boldsymbol{B} = -\int_\Omega \sum_\alpha \sum_j \frac{\partial}{\partial x_j}(\phi_\varepsilon(\boldsymbol{h}(\boldsymbol{x},t)))(\operatorname{cof} D_1(\boldsymbol{h}(\boldsymbol{x},t)))_{\alpha j} h_{\alpha,t} dx +$$

$$-\int_\Omega \sum_\alpha \sum_j \phi_\varepsilon(\boldsymbol{h}(\boldsymbol{x},t))(\operatorname{cof} D(\boldsymbol{h}(\boldsymbol{x},t)))_{\alpha j,j} h_{\alpha,t} dx$$

The second term equals zero by Lemma 15.1.8. Simplifying the first term yields

$$\begin{aligned}\boldsymbol{B} &= -\int_\Omega \sum_\alpha \sum_j \sum_\beta \phi_{\varepsilon,\beta}(\boldsymbol{h}(\boldsymbol{x},t)) h_{\beta,j} h_{\alpha,t} (\operatorname{cof} D_1(\boldsymbol{h}(\boldsymbol{x},t)))_{\alpha j} dx \\ &= -\int_\Omega \sum_\alpha \sum_\beta \phi_{\varepsilon,\beta}(\boldsymbol{h}(\boldsymbol{x},t)) h_{\alpha,t} \sum_j h_{\beta,j} (\operatorname{cof} D_1(\boldsymbol{h}(\boldsymbol{x},t)))_{\alpha j} dx\end{aligned}$$

Now the sum on j is the dot product of the β^{th} row with the α^{th} row of the cofactor matrix which equals zero unless $\beta = \alpha$ because it would be a cofactor expansion of a matrix with two equal rows. When $\beta = \alpha$, the sum on j reduces to $\det(D_1(\boldsymbol{h}(\boldsymbol{x},t)))$. Thus $\boldsymbol{B}$ reduces to

$$= -\int_\Omega \sum_\alpha \phi_{\varepsilon,\alpha}(\boldsymbol{h}(\boldsymbol{x},t)) h_{\alpha,t} \det(D_1(\boldsymbol{h}(\boldsymbol{x},t)))\, dx$$

Which is the same thing as $\boldsymbol{A}$, but with the opposite sign. Hence $\boldsymbol{A}+\boldsymbol{B}$ in 15.3 is 0 and $H'(t) = 0$ and so H is a constant on $[a,b]$.

Finally consider the last claim. If $\boldsymbol{g},\hat{\boldsymbol{g}}$ both work in the definition for the degree, then consider $\boldsymbol{h}(\boldsymbol{x},t) \equiv t\boldsymbol{g}(\boldsymbol{x}) + (1-t)\hat{\boldsymbol{g}}(\boldsymbol{x}) - \boldsymbol{y}$ for $t \in [0,1]$. For $\boldsymbol{x} \in \partial\Omega$,

$$\begin{aligned}&|t\boldsymbol{g}(\boldsymbol{x}) + (1-t)\hat{\boldsymbol{g}}(\boldsymbol{x}) - \boldsymbol{y}| \\ = \;& |t(\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x})) + (1-t)(\hat{\boldsymbol{g}}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x})) + \boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{y}|\end{aligned}$$

$$\begin{aligned}\geq \;& |\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{y}| - |t(\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x})) + (1-t)(\hat{\boldsymbol{g}}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}))| \\ \geq \;& \operatorname{dist}(\boldsymbol{f}(\partial\Omega), \boldsymbol{y}) - (t\|\boldsymbol{g} - \boldsymbol{f}\|_\infty + (1-t)\|\hat{\boldsymbol{g}} - \boldsymbol{f}\|_\infty) \\ > \;& \operatorname{dist}(\boldsymbol{f}(\partial\Omega), \boldsymbol{y}) - (t\delta + (1-t)\delta) = 0\end{aligned}$$

From Lemma 15.1.12, $\boldsymbol{h}$ satisfies what is needed for the first part of this lemma. Namely, $\boldsymbol{0} \notin \boldsymbol{h}(\partial\Omega \times [0,1])$. Then from the first part of this lemma, if

$$0 < \varepsilon < \operatorname{dist}(\boldsymbol{0}, \boldsymbol{h}(\partial\Omega \times [0,1]))$$

is also sufficiently small that the second and last equations hold in what follows,

$$\begin{aligned} d(\boldsymbol{f},\Omega,\boldsymbol{y}) &= \sum\left\{\operatorname{sgn}(\det(D\boldsymbol{g}(\boldsymbol{x}))) : \boldsymbol{x}\in \boldsymbol{g}^{-1}(\boldsymbol{y})\right\} \\ &= \int_\Omega \phi_\varepsilon(\boldsymbol{h}(\boldsymbol{x},1))\det D_1\boldsymbol{h}(\boldsymbol{x},1)\,dx \\ &= \int_\Omega \phi_\varepsilon(\boldsymbol{h}(\boldsymbol{x},0))\det D_1\boldsymbol{h}(\boldsymbol{x},0)\,dx \\ &= \sum\left\{\operatorname{sgn}(\det(D\hat{\boldsymbol{g}}(\boldsymbol{x}))) : \boldsymbol{x}\in \hat{\boldsymbol{g}}^{-1}(\boldsymbol{y})\right\} \blacksquare \end{aligned}$$

15.2 Properties of the Degree

Now that the degree for a continuous function has been defined, it is time to consider properties of the degree. In particular, it is desirable to prove a theorem about homotopy invariance which depends only on continuity considerations.

Theorem 15.2.1 *If $\boldsymbol{h}$ is in $C\left(\overline{\Omega}\times[a,b],\mathbb{R}^p\right)$, and $\boldsymbol{0}\notin \boldsymbol{h}(\partial\Omega\times[a,b])$ for each t, then $t\to d(\boldsymbol{h}(\cdot,t),\Omega,\boldsymbol{0})$ is constant for $t\in[a,b]$.*

Proof: Let $0<\delta=\min|\boldsymbol{h}(\partial\Omega\times[a,b])|$. By Corollary 15.1.1, there exists

$$\boldsymbol{h}_m(\cdot,t)=\sum_{k=0}^m p_k(t)\,\boldsymbol{h}(\cdot,t_k)$$

for $p_k(t)$ some polynomial in t of degree m such that

$$\max_{t\in[a,b]}\|\boldsymbol{h}_m(\cdot,t)-\boldsymbol{h}(\cdot,t)\|_{\infty,\overline{\Omega}}<\delta \tag{15.4}$$

Now replace $\boldsymbol{h}(\cdot,t_k)$ with $\boldsymbol{g}_k^n(\cdot)\in C^\infty\left(\overline{\Omega},\mathbb{R}^p\right)$ and let

$$\boldsymbol{g}_n(\cdot,t)\equiv\sum_{k=0}^m p_k(t)\,\boldsymbol{g}_k^n(\cdot)$$

where this function $\boldsymbol{g}_k^n$ is close enough to $\boldsymbol{h}(\cdot,t_k)$ that

$$\max_{t\in[a,b]}\|\boldsymbol{g}_n(\cdot,t)-\boldsymbol{h}(\cdot,t)\|_{\infty,\overline{\Omega}}<\delta \tag{15.5}$$

Now $\boldsymbol{g}_n\in C^\infty\left(\overline{\Omega}\times[a,b];\mathbb{R}^p\right)$ because all partial derivatives with respect to either t or $\boldsymbol{x}$ are continuous.

Let $\tau\in(a,b]$. Let $\boldsymbol{g}_{a\tau}(\boldsymbol{x},t)\equiv\boldsymbol{g}_n(\boldsymbol{x},t)-\left(\frac{\tau-t}{\tau-a}\boldsymbol{y}_a+\boldsymbol{y}_\tau\frac{t-a}{\tau-a}\right)$ where $\boldsymbol{y}_a$ is a regular value of $\boldsymbol{g}_n(\cdot,a)$, $\boldsymbol{y}_\tau$ is a regular value of $\boldsymbol{g}_n(\boldsymbol{x},\tau)$ and both $\boldsymbol{y}_a,\boldsymbol{y}_\tau$ are so small that

$$\max_{t\in[a,b]}\|\boldsymbol{g}_{a\tau}(\cdot,t)-\boldsymbol{h}(\cdot,t)\|_{\infty,\overline{\Omega}}<\delta. \tag{15.6}$$

This uses Lemma 15.1.3. Thus if $\boldsymbol{g}_{a\tau}(\boldsymbol{x},\tau)=\boldsymbol{0}$, then $\boldsymbol{g}_n(\boldsymbol{x},\tau)-\boldsymbol{y}_\tau=\boldsymbol{0}$ so $\boldsymbol{0}$ is a regular value for $\boldsymbol{g}_{a\tau}(\cdot,\tau)$. If $\boldsymbol{g}_{a\tau}(\boldsymbol{x},a)=\boldsymbol{0}$, then $\boldsymbol{g}_n(\boldsymbol{x},t)=\boldsymbol{y}_a$ so $\boldsymbol{0}$ is also a regular value for $\boldsymbol{g}_{a\tau}(\cdot,a)$. From 15.6, $\operatorname{dist}(\boldsymbol{g}_{a\tau}(\partial\Omega\times[a,b]),\boldsymbol{0})>0$ as in Lemma 15.1.5. Choosing $\varepsilon<$

$\mathrm{dist}\left(\boldsymbol{g}_{a\tau}\left(\partial\Omega\times[a,b]\right),\mathbf{0}\right)$, it follows from Lemma 15.1.13, the definition of the degree, and Lemma 15.1.11 in the first and last equations that for ε small enough,

$$\begin{aligned} d\left(\boldsymbol{h}\left(\cdot,a\right),\Omega,\mathbf{0}\right) &= \int_{\Omega}\phi_{\varepsilon}\left(\boldsymbol{g}_{a\tau}\left(\boldsymbol{x},a\right)\right)\det D_{1}\boldsymbol{g}_{a\tau}\left(\boldsymbol{x},a\right)dx \\ &= \int_{\Omega}\phi_{\varepsilon}\left(\boldsymbol{g}_{a\tau}\left(\boldsymbol{x},\tau\right)\right)\det D_{1}\boldsymbol{g}_{a\tau}\left(\boldsymbol{x},\tau\right)dx = d\left(\boldsymbol{h}\left(\cdot,\tau\right),\Omega,\mathbf{0}\right) \end{aligned}$$

Since τ is arbitrary, this proves the theorem. ■

Now the following theorem is a summary of the main result on properties of the degree.

Theorem 15.2.2 *Definition 15.1.6 is well defined and the degree satisfies the following properties.*

1. *(homotopy invariance) If* $\boldsymbol{h}\in C\left(\overline{\Omega}\times[0,1],\mathbb{R}^{p}\right)$ *and* $\boldsymbol{y}\left(t\right)\notin\boldsymbol{h}\left(\partial\Omega,t\right)$ *for all* $t\in[0,1]$ *where* $\boldsymbol{y}$ *is continuous, then*

$$t\to d\left(\boldsymbol{h}\left(\cdot,t\right),\Omega,\boldsymbol{y}\left(t\right)\right)$$

   is constant for $t\in[0,1]$.

2. *If* $\Omega\supseteq\Omega_{1}\cup\Omega_{2}$ *where* $\Omega_{1}\cap\Omega_{2}=\emptyset$, *for* Ω_{i} *an open set, then if*

$$\boldsymbol{y}\notin\boldsymbol{f}\left(\overline{\Omega}\setminus\left(\Omega_{1}\cup\Omega_{2}\right)\right),$$

   then

$$d\left(\boldsymbol{f},\Omega_{1},\boldsymbol{y}\right)+d\left(\boldsymbol{f},\Omega_{2},\boldsymbol{y}\right)=d\left(\boldsymbol{f},\Omega,\boldsymbol{y}\right)$$

3. $d\left(I,\Omega,\boldsymbol{y}\right)=1$ *if* $\boldsymbol{y}\in\Omega$.

4. $d\left(\boldsymbol{f},\Omega,\cdot\right)$ *is continuous and constant on every connected component of* $\mathbb{R}^{p}\setminus\boldsymbol{f}\left(\partial\Omega\right)$.

5. $d\left(\boldsymbol{g},\Omega,\boldsymbol{y}\right)=d\left(\boldsymbol{f},\Omega,\boldsymbol{y}\right)$ *if* $\boldsymbol{g}|_{\partial\Omega}=\boldsymbol{f}|_{\partial\Omega}$.

6. *If* $\boldsymbol{y}\notin\boldsymbol{f}\left(\partial\Omega\right)$, *and if* $d\left(\boldsymbol{f},\Omega,\boldsymbol{y}\right)\neq0$, *then there exists* $\boldsymbol{x}\in\Omega$ *such that* $\boldsymbol{f}\left(\boldsymbol{x}\right)=\boldsymbol{y}$.

Proof: That the degree is well defined follows from Lemma 15.1.13.

Consider 1., the first property about homotopy. This follows from Theorem 15.2.1 applied to $H\left(\boldsymbol{x},t\right)\equiv\boldsymbol{h}\left(\boldsymbol{x},t\right)-\boldsymbol{y}\left(t\right)$.

Consider 2. where $\boldsymbol{y}\notin\boldsymbol{f}\left(\overline{\Omega}\setminus\left(\Omega_{1}\cup\Omega_{2}\right)\right)$. Note that

$$\mathrm{dist}\left(\boldsymbol{y},\boldsymbol{f}\left(\overline{\Omega}\setminus\left(\Omega_{1}\cup\Omega_{2}\right)\right)\right)\leq\mathrm{dist}\left(\boldsymbol{y},\boldsymbol{f}\left(\partial\Omega\right)\right)$$

Then let $\boldsymbol{g}$ be in $C\left(\overline{\Omega};\mathbb{R}^{p}\right)$ and

$$\begin{aligned} \left\|\boldsymbol{g}-\boldsymbol{f}\right\|_{\infty} &< \mathrm{dist}\left(\boldsymbol{y},\boldsymbol{f}\left(\overline{\Omega}\setminus\left(\Omega_{1}\cup\Omega_{2}\right)\right)\right) \\ &\leq \min\left(\mathrm{dist}\left(\boldsymbol{y},\boldsymbol{f}\left(\partial\Omega_{1}\right)\right),\mathrm{dist}\left(\boldsymbol{y},\boldsymbol{f}\left(\partial\Omega_{2}\right)\right),\mathrm{dist}\left(\boldsymbol{y},\boldsymbol{f}\left(\partial\Omega\right)\right)\right) \end{aligned}$$

where $\boldsymbol{y}$ is a regular value of $\boldsymbol{g}$. Then by definition,

$$d\left(\boldsymbol{f},\Omega,\boldsymbol{y}\right)\equiv\sum\left\{\det\left(D\boldsymbol{g}\left(\boldsymbol{x}\right)\right):\boldsymbol{x}\in\boldsymbol{g}^{-1}\left(\boldsymbol{y}\right)\right\}$$

$$\begin{aligned} &= \sum \left\{ \det(D\boldsymbol{g}(\boldsymbol{x})) : \boldsymbol{x} \in \boldsymbol{g}^{-1}(\boldsymbol{y}), \boldsymbol{x} \in \Omega_1 \right\} \\ &\quad + \sum \left\{ \det(D\boldsymbol{g}(\boldsymbol{x})) : \boldsymbol{x} \in \boldsymbol{g}^{-1}(\boldsymbol{y}), \boldsymbol{x} \in \Omega_2 \right\} \\ &\equiv d(\boldsymbol{f}, \Omega_1, \boldsymbol{y}) + d(\boldsymbol{f}, \Omega_2, \boldsymbol{y}) \end{aligned}$$

It is of course obvious that this can be extended by induction to any finite number of disjoint open sets Ω_i.

Note that 3. is obvious because $I(\boldsymbol{x}) = \boldsymbol{x}$ and so if $\boldsymbol{y} \in \Omega$, then $I^{-1}(\boldsymbol{y}) = \boldsymbol{y}$ and $DI(\boldsymbol{x}) = I$ for any $\boldsymbol{x}$ so the definition gives 3.

Now consider 4. Let U be a connected component of $\mathbb{R}^p \setminus \boldsymbol{f}(\partial\Omega)$. This is open as well as connected and arc wise connected by Theorem 3.11.12. Hence, if $\boldsymbol{u}, \boldsymbol{v} \in U$, there is a continuous function $\boldsymbol{y}(t)$ which is in U such that $\boldsymbol{y}(0) = \boldsymbol{u}$ and $\boldsymbol{y}(1) = \boldsymbol{v}$. By homotopy invariance, it follows $d(\boldsymbol{f}, \Omega, \boldsymbol{y}(t))$ is constant. Thus $d(\boldsymbol{f}, \Omega, \boldsymbol{u}) = d(\boldsymbol{f}, \Omega, \boldsymbol{v})$.

Next consider 5. When $\boldsymbol{f} = \boldsymbol{g}$ on $\partial\Omega$, it follows that if $\boldsymbol{y} \notin \boldsymbol{f}(\partial\Omega)$, then $\boldsymbol{y} \notin \boldsymbol{f}(\boldsymbol{x}) + t(\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}))$ for $t \in [0,1]$ and $\boldsymbol{x} \in \partial\Omega$ so

$$d(\boldsymbol{f} + t(\boldsymbol{g} - \boldsymbol{f}), \Omega, \boldsymbol{y})$$

is constant for $t \in [0,1]$ by homotopy invariance in part 1. Therefore, let $t = 0$ and then $t = 1$ to obtain 5.

Claim 6. follows from Lemma 15.1.13 which says that if $\boldsymbol{y} \notin \boldsymbol{f}(\overline{\Omega})$, then $d(\boldsymbol{f}, \Omega, \boldsymbol{y}) = 0$. ■

From the above, there is an easy corollary which gives related properties of the degree.

Corollary 15.2.3 *The following additional properties of the degree are also valid.*

1. *If $\boldsymbol{y} \notin \boldsymbol{f}(\overline{\Omega} \setminus \Omega_1)$ and Ω_1 is an open subset of Ω, then $d(\boldsymbol{f}, \Omega, \boldsymbol{y}) = d(\boldsymbol{f}, \Omega_1, \boldsymbol{y})$.*
2. *$d(\cdot, \Omega, \boldsymbol{y})$ is defined and constant on*

$$\left\{ \boldsymbol{g} \in C(\overline{\Omega}; \mathbb{R}^p) : \|\boldsymbol{g} - \boldsymbol{f}\|_\infty < r \right\}$$

   where $r = \operatorname{dist}(\boldsymbol{y}, \boldsymbol{f}(\partial\Omega))$.

3. *If $\boldsymbol{y} \in \boldsymbol{f}(\Omega)$, $\operatorname{dist}(\boldsymbol{y}, \boldsymbol{f}(\partial\Omega)) \geq \delta$ and $|\boldsymbol{z} - \boldsymbol{y}| < \delta$, then $d(\boldsymbol{f}, \Omega, \boldsymbol{y}) = d(\boldsymbol{f}, \Omega, \boldsymbol{z})$.*

Proof: Consider 1. You can take $\Omega_2 = \emptyset$ in 2 of Theorem 15.2.2 or you can modify the proof of 2 slightly. Consider 2. To verify, let $\boldsymbol{h}(\boldsymbol{x}, t) = \boldsymbol{f}(\boldsymbol{x}) + t(\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}))$. Then note that $\boldsymbol{y} \notin \boldsymbol{h}(\partial\Omega, t)$ and use Property 1 of Theorem 15.2.2. Finally, consider 3. Let $\boldsymbol{y}(t) \equiv (1-t)\boldsymbol{y} + t\boldsymbol{z}$. Then for $\boldsymbol{x} \in \partial\Omega$

$$\begin{aligned} |(1-t)\boldsymbol{y} + t\boldsymbol{z} - \boldsymbol{f}(\boldsymbol{x})| &= |\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}) + t(\boldsymbol{z} - \boldsymbol{y})| \\ &\geq \delta - t|\boldsymbol{z} - \boldsymbol{y}| > \delta - \delta = 0 \end{aligned}$$

Then by 1 of Theorem 15.2.2, $d(\boldsymbol{f}, \Omega, (1-t)\boldsymbol{y} + t\boldsymbol{z})$ is constant. When $t = 0$ you get $d(\boldsymbol{f}, \Omega, \boldsymbol{y})$ and when $t = 1$ you get $d(\boldsymbol{f}, \Omega, \boldsymbol{z})$. ■

15.3 Brouwer Fixed Point Theorem

The degree makes it possible to give a very simple proof of the Brouwer fixed point theorem.

Theorem 15.3.1 *(Brouwer fixed point) Let $B = \overline{B(\mathbf{0},r)} \subseteq \mathbb{R}^p$ and let $\boldsymbol{f} : B \to B$ be continuous. Then there exists a point $\boldsymbol{x} \in B$, such that $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{x}$.*

Proof: Assume there is no fixed point. Consider $\boldsymbol{h}(\boldsymbol{x},t) \equiv t\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{x}$ for $t \in [0,1]$. Then for $\|\boldsymbol{x}\| = r$,

$$\mathbf{0} \notin t\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{x},\ t \in [0,1]$$

By homotopy invariance, $t \to d(t\boldsymbol{f} - I, B, \mathbf{0})$ is constant. But when $t = 0$, this is

$$d(-I, B, \mathbf{0}) = (-1)^n \neq 0.$$

Hence $d(\boldsymbol{f} - I, B, \mathbf{0}) \neq 0$ so there exists $\boldsymbol{x}$ such that $\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{x} = \mathbf{0}$. ■

You can use standard stuff from Hilbert space to get this the fixed point theorem for a compact convex set. Let K be a closed bounded convex set and let $\boldsymbol{f} : K \to K$ be continuous. Let P be the projection map onto K as in Problem 10 on Page 152. Then P is continuous because $|P\boldsymbol{x} - P\boldsymbol{y}| \leq |\boldsymbol{x} - \boldsymbol{y}|$. Recall why this is. From the characterization of the projection map P, $(\boldsymbol{x} - P\boldsymbol{x}, \boldsymbol{y} - P\boldsymbol{x}) \leq 0$ for all $\boldsymbol{y} \in K$. Therefore,

$$(\boldsymbol{x} - P\boldsymbol{x}, P\boldsymbol{y} - P\boldsymbol{x}) \leq 0,\ (\boldsymbol{y} - P\boldsymbol{y}, P\boldsymbol{x} - P\boldsymbol{y}) \leq 0 \text{ so } (\boldsymbol{y} - P\boldsymbol{y}, P\boldsymbol{y} - P\boldsymbol{x}) \geq 0$$

Hence, subtracting the first from the last,

$$(\boldsymbol{y} - P\boldsymbol{y} - (\boldsymbol{x} - P\boldsymbol{x}), P\boldsymbol{y} - P\boldsymbol{x}) \geq 0$$

consequently,

$$|\boldsymbol{x} - \boldsymbol{y}|\,|P\boldsymbol{y} - P\boldsymbol{x}| \geq (\boldsymbol{y} - \boldsymbol{x}, P\boldsymbol{y} - P\boldsymbol{x}) \geq |P\boldsymbol{y} - P\boldsymbol{x}|^2$$

and so $|P\boldsymbol{y} - P\boldsymbol{x}| \leq |\boldsymbol{y} - \boldsymbol{x}|$ as claimed.

Now let r be so large that $K \subseteq B(\mathbf{0},r)$. Then consider $\boldsymbol{f} \circ P$. This map takes $\overline{B(\mathbf{0},r)} \to \overline{B(\mathbf{0},r)}$. In fact it maps $\overline{B(\mathbf{0},r)}$ to K. Therefore, being the composition of continuous functions, it is continuous and so has a fixed point in $\overline{B(\mathbf{0},r)}$ denoted as $\boldsymbol{x}$. Hence $\boldsymbol{f}(P(\boldsymbol{x})) = \boldsymbol{x}$. Now, since $\boldsymbol{f}$ maps into K, it follows that $\boldsymbol{x} \in K$. Hence $P\boldsymbol{x} = \boldsymbol{x}$ and so $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{x}$. This has proved the following general Brouwer fixed point theorem.

Theorem 15.3.2 *Let $\boldsymbol{f} : K \to K$ be continuous where K is compact and convex and nonempty, $K \subseteq \mathbb{R}^p$. Then $\boldsymbol{f}$ has a fixed point.*

Definition 15.3.3 *$\boldsymbol{f}$ is a retract of $\overline{B(\mathbf{0},r)}$ onto $\partial B(0,r)$ if $\boldsymbol{f}$ is continuous,*

$$\boldsymbol{f}\left(\overline{B(\mathbf{0},r)}\right) \subseteq \partial B(\mathbf{0},r)$$

and $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{x}$ for all $\boldsymbol{x} \in \partial B(\mathbf{0},r)$.

Theorem 15.3.4 *There does not exist a retract of $\overline{B(\mathbf{0},r)}$ onto $\partial B(\mathbf{0},r)$, its boundary.*

Proof: Suppose $\boldsymbol{f}$ were such a retract. Then for all $\boldsymbol{x} \in \partial B(\mathbf{0},r)$, $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{x}$ and so from the properties of the degree, the one which says if two functions agree on $\partial\Omega$, then they have the same degree,

$$1 = d(I, B(\mathbf{0},r), \mathbf{0}) = d(\boldsymbol{f}, B(\mathbf{0},r), \mathbf{0})$$

which is clearly impossible because $\boldsymbol{f}^{-1}(\mathbf{0}) = \emptyset$ which implies $d(\boldsymbol{f}, B(\mathbf{0},r), \mathbf{0}) = 0$. ■

You should now use this theorem to give another proof of the Brouwer fixed point theorem.

15.4 Borsuk's Theorem

In this section is an important theorem which can be used to verify that $d(\boldsymbol{f},\Omega,\boldsymbol{y}) \neq 0$. This is significant because when this is known, it follows from Theorem 15.2.2 that $\boldsymbol{f}^{-1}(\boldsymbol{y}) \neq \emptyset$. In other words there exists $\boldsymbol{x} \in \Omega$ such that $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{y}$.

Definition 15.4.1 *A bounded open set, Ω is symmetric if $-\Omega = \Omega$. A continuous function, $\boldsymbol{f} : \overline{\Omega} \to \mathbb{R}^p$ is odd if $\boldsymbol{f}(-\boldsymbol{x}) = -\boldsymbol{f}(\boldsymbol{x})$.*

Suppose Ω is symmetric and $\boldsymbol{g} \in C^\infty\left(\overline{\Omega};\mathbb{R}^p\right)$ is an odd map for which $\mathbf{0}$ is a regular value. Then the chain rule implies $D\boldsymbol{g}(-\boldsymbol{x}) = D\boldsymbol{g}(\boldsymbol{x})$ and so $d(\boldsymbol{g},\Omega,\mathbf{0})$ must equal an odd integer because if $\boldsymbol{x} \in \boldsymbol{g}^{-1}(\mathbf{0})$, it follows that $-\boldsymbol{x} \in \boldsymbol{g}^{-1}(\mathbf{0})$ also and since $D\boldsymbol{g}(-\boldsymbol{x}) = D\boldsymbol{g}(\boldsymbol{x})$, it follows the overall contribution to the degree from $\boldsymbol{x}$ and $-\boldsymbol{x}$ must be an even integer. Also $\mathbf{0} \in \boldsymbol{g}^{-1}(\mathbf{0})$ and so the degree equals an even integer added to $\operatorname{sgn}(\det D\boldsymbol{g}(\mathbf{0}))$, an odd integer, either -1 or 1. It seems reasonable to expect that something like this would hold for an arbitrary continuous odd function defined on symmetric Ω. In fact this is the case and this is next. The following lemma is the key result used. This approach is due to Gromes [24]. See also Deimling [12] which is where I found this argument. I think it is one of the cleverest calculus manipulations I have seen.

The idea is to start with a smooth odd map and approximate it with a smooth odd map which also has $\mathbf{0}$ a regular value. Note that $\mathbf{0}$ is achieved because $\boldsymbol{g}(\mathbf{0}) = -\boldsymbol{g}(\mathbf{0})$.

Process: Suppose $\boldsymbol{h}_0 \in C^\infty\left(\overline{\Omega},\mathbb{R}^p\right)$ is odd and infinitely differentiable and

$$\det(D\boldsymbol{h}_0(\mathbf{0})) \neq 0.$$

Let Ω_k be those points of Ω where $x_k \neq 0$. Here $\boldsymbol{x} \equiv (x_1,...,x_p)$. Then $\boldsymbol{x} \to \frac{\boldsymbol{h}_0(\boldsymbol{x})}{x_k^3}$ is a smooth map defined on Ω_k so by Sard's theorem, its singular values do not contain $B(\mathbf{0},\eta)$. Therefore, there is $\boldsymbol{y}^k$ with $\boldsymbol{y}^k$ a regular value and $\left\|\boldsymbol{y}^k\right\| < \eta$ where $\eta > 0$ is given. Then consider $\hat{\boldsymbol{h}}(\boldsymbol{x}) \equiv \boldsymbol{h}_0(\boldsymbol{x}) - x_k^3\boldsymbol{y}^k$. I want to argue that $\mathbf{0}$ is a regular value of $\hat{\boldsymbol{h}}$ on Ω_k. Note that $\frac{\boldsymbol{h}_0(\boldsymbol{x})}{x_k^3} = \boldsymbol{y}^k$ if and only if $\hat{\boldsymbol{h}}(\boldsymbol{x}) = \mathbf{0}$.

Letting $\boldsymbol{f}(\boldsymbol{x}) \equiv \frac{\boldsymbol{h}_0(\boldsymbol{x})}{x_k^3}, \hat{\boldsymbol{h}}(\boldsymbol{x}) = x_k^3\left(\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{y}^k\right)$ and $D\boldsymbol{f}(\boldsymbol{x})$ is invertible at the $\boldsymbol{x}$ of interest, one where $\hat{\boldsymbol{h}}(\boldsymbol{x}) = \mathbf{0}$. Then if $r_k(\boldsymbol{x}) = x_k^3$, it follows from the definition of the derivative that

$$D\hat{\boldsymbol{h}}(\boldsymbol{x})(\boldsymbol{u}) = D(r_k(\boldsymbol{x}))(\boldsymbol{u})\left(\overset{=\mathbf{0}}{\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{y}^k}\right) + x_k^3 D\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{u}). \tag{15.7}$$

At the point of interest, the first term is $\mathbf{0}$ and so

$$\det\left(D\hat{\boldsymbol{h}}(\boldsymbol{x})\right) = x_k^3 \det(D\boldsymbol{f}(\boldsymbol{x})) = \det\left(x_k^3 D\boldsymbol{f}(\boldsymbol{x})\right) \neq 0.$$

If whenever $\boldsymbol{h}_0(\boldsymbol{x}) = \mathbf{0}, \det(D\boldsymbol{h}_0(\boldsymbol{x})) \neq 0$, for $\boldsymbol{x} \in \mathscr{U} \subseteq \Omega$, does this continue to hold for $\hat{\boldsymbol{h}}(\boldsymbol{x})$ for points of $\mathscr{U}$? The only points of concern are those for which $x_k = 0$ and for these points $\hat{\boldsymbol{h}}(\boldsymbol{x}) = \boldsymbol{h}_0(\boldsymbol{x})$ and $D\hat{\boldsymbol{h}}(\boldsymbol{x}) = D\boldsymbol{h}_0(\boldsymbol{x})$ because $3x_k^2 = 0$ when $x_k = 0$. Thus the new function $\hat{\boldsymbol{h}}$ has $\mathbf{0}$ a regular value for all $\boldsymbol{x} \in \mathscr{U} \cup \Omega_k$. This **Process** is the basis for the following lemma.

Lemma 15.4.2 *Let $\boldsymbol{h}_0 \in C^\infty\left(\overline{\Omega}, \mathbb{R}^p\right)$ is odd and infinitely differentiable and*

$$\det\left(D\boldsymbol{h}_0(\boldsymbol{0})\right) \neq 0$$

for Ω a symmetric open set and let $\eta > 0$. Then there are vectors $\boldsymbol{y}^k$ each with $\left\|\boldsymbol{y}^k\right\| < \eta$ such that $\boldsymbol{h}(\boldsymbol{x}) \equiv \boldsymbol{h}_0(\boldsymbol{x}) - \sum_{k=1}^p x_k^3 \boldsymbol{y}^k$ has $\boldsymbol{0}$ as a regular value.

Proof: Use the above process leading to 15.7 repeatedly. Start with $\boldsymbol{h}_0$ which has $\boldsymbol{0}$ a regular value on $\{\boldsymbol{0}\}$. Then use the process to get $\boldsymbol{h}_1(\boldsymbol{x}) = \boldsymbol{h}_0(\boldsymbol{x}) - \boldsymbol{y}^1 x_1^3$ which has $\boldsymbol{0}$ as a regular value on $\{\boldsymbol{0}\} \cup \Omega_1$. Then repeat the process to get $\boldsymbol{h}_2(\boldsymbol{x}) = \boldsymbol{h}_1(\boldsymbol{x}) - \boldsymbol{y}^2 x_2^3$ which has $\boldsymbol{0}$ as a regular value on $\{\boldsymbol{0}\} \cup \Omega_1 \cup \Omega_2$. Continue this way and let $\boldsymbol{h} = \boldsymbol{h}_p$ which has $\boldsymbol{0}$ a regular value on $\{\boldsymbol{0}\} \cup \Omega_1 \cup \cdots \cup \Omega_p = \Omega$. ■

Lemma 15.4.3 *Let $\boldsymbol{g} \in C^\infty\left(\overline{\Omega}; \mathbb{R}^p\right)$ be an odd map. Then for every $\varepsilon > 0$, there exists $\boldsymbol{h} \in C^\infty\left(\overline{\Omega}; \mathbb{R}^p\right)$ such that $\boldsymbol{h}$ is also an odd map, $\|\boldsymbol{h} - \boldsymbol{g}\|_\infty < \varepsilon$, and $\boldsymbol{0}$ is a regular value of $\boldsymbol{h}, \boldsymbol{0} \notin \boldsymbol{g}(\partial\Omega)$. Here Ω is a symmetric bounded open set. In addition, $d(\boldsymbol{g}, \Omega, \boldsymbol{0})$ is an odd integer.*

Proof: In this argument $\eta > 0$ will be a small positive number. Let $\boldsymbol{h}_0(\boldsymbol{x}) = \boldsymbol{g}(\boldsymbol{x}) + \eta\boldsymbol{x}$ where η is chosen such that $\det D\boldsymbol{h}_0(\boldsymbol{0}) \neq 0$. See Lemma 8.10.2. Note that $\boldsymbol{h}_0$ is odd and $\boldsymbol{0}$ is a value of $\boldsymbol{h}_0$ thanks to $\boldsymbol{h}_0(\boldsymbol{0}) = \boldsymbol{0}$. This has taken care of $\boldsymbol{0}$. However, it is not known whether $\boldsymbol{0}$ is a regular value of $\boldsymbol{h}_0$ because there may be other $\boldsymbol{x}$ where $\boldsymbol{h}_0(\boldsymbol{x}) = \boldsymbol{0}$. By Lemma 15.4.2, there are vectors $\boldsymbol{y}^j$ with $\left\|\boldsymbol{y}^k\right\| \leq \eta$ and $\boldsymbol{0}$ is a regular value of $\boldsymbol{h}(\boldsymbol{x}) \equiv \boldsymbol{h}_0(\boldsymbol{x}) - \sum_{j=1}^p \boldsymbol{y}^j x_j^3$. Then

$$\begin{aligned} \|\boldsymbol{h} - \boldsymbol{g}\|_{\infty,\,\overline{\Omega}} &\leq \max_{\boldsymbol{x}\in\Omega}\left\{\|\eta\boldsymbol{x}\| + \sum_{k=1}^p \left\|\boldsymbol{y}^k\right\| \|\boldsymbol{x}\|\right\} \\ &\leq \eta\left((p+1)\operatorname{diam}(\Omega)\right) < \varepsilon < \operatorname{dist}\left(\boldsymbol{g}(\partial\Omega), \boldsymbol{0}\right) \end{aligned}$$

provided η was chosen sufficiently small to begin with.

So what is $d(\boldsymbol{h}, \Omega, \boldsymbol{0})$? Since $\boldsymbol{0}$ is a regular value and $\boldsymbol{h}$ is odd,

$$\boldsymbol{h}^{-1}(\boldsymbol{0}) = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_r, -\boldsymbol{x}_1, \cdots, -\boldsymbol{x}_r, \boldsymbol{0}\}.$$

So consider $D\boldsymbol{h}(\boldsymbol{x})$ and $D\boldsymbol{h}(-\boldsymbol{x})$.

$$\begin{aligned} D\boldsymbol{h}(-\boldsymbol{x})\boldsymbol{u} + \boldsymbol{o}(\boldsymbol{u}) &= \boldsymbol{h}(-\boldsymbol{x} + \boldsymbol{u}) - \boldsymbol{h}(-\boldsymbol{x}) = -\boldsymbol{h}(\boldsymbol{x} + (-\boldsymbol{u})) + \boldsymbol{h}(\boldsymbol{x}) \\ &= -\left(D\boldsymbol{h}(\boldsymbol{x})(-\boldsymbol{u})\right) + \boldsymbol{o}(-\boldsymbol{u}) = D\boldsymbol{h}(\boldsymbol{x})(\boldsymbol{u}) + \boldsymbol{o}(\boldsymbol{u}) \end{aligned}$$

Hence $D\boldsymbol{h}(\boldsymbol{x}) = D\boldsymbol{h}(-\boldsymbol{x})$ and so the determinants of these two are the same. It follows from the definition that $d(\boldsymbol{g}, \Omega, \boldsymbol{0}) = d(\boldsymbol{h}, \Omega, \boldsymbol{0})$

$$\begin{aligned} &= \sum_{i=1}^r \operatorname{sgn}\left(\det\left(D\boldsymbol{h}(\boldsymbol{x}_i)\right)\right) + \sum_{i=1}^r \operatorname{sgn}\left(\det\left(D\boldsymbol{h}(-\boldsymbol{x}_i) + \operatorname{sgn}\left(\det\left(D\boldsymbol{h}(\boldsymbol{0})\right)\right)\right)\right) \\ &= 2m \pm 1 \text{ some integer } m \end{aligned}$$

■

Theorem 15.4.4 *(Borsuk) Let $\boldsymbol{f} \in C\left(\overline{\Omega}; \mathbb{R}^p\right)$ be odd and let Ω be symmetric with $\boldsymbol{0} \notin \boldsymbol{f}(\partial\Omega)$. Then $d(\boldsymbol{f}, \Omega, \boldsymbol{0})$ equals an odd integer.*

Proof: Let ψ_n be a mollifier which is symmetric, $\psi(-\boldsymbol{x}) = \psi(\boldsymbol{x})$. Also recall that $\boldsymbol{f}$ is the restriction to $\overline{\Omega}$ of a continuous function, still denoted as $\boldsymbol{f}$ which is defined on all of $\mathbb{R}^p$. Let $\boldsymbol{g}$ be the odd part of this function. That is,

$$\boldsymbol{g}(\boldsymbol{x}) \equiv \frac{1}{2}(\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(-\boldsymbol{x}))$$

Since $\boldsymbol{f}$ is odd, $\boldsymbol{g} = \boldsymbol{f}$ on $\overline{\Omega}$. Then

$$\boldsymbol{g}_n(-\boldsymbol{x}) \equiv \boldsymbol{g} * \psi_n(-\boldsymbol{x}) = \int_\Omega \boldsymbol{g}(-\boldsymbol{x}-\boldsymbol{y})\,\psi_n(\boldsymbol{y})\,dy$$

$$= -\int_\Omega \boldsymbol{g}(\boldsymbol{x}+\boldsymbol{y})\,\psi_n(\boldsymbol{y})\,dy = -\int_\Omega \boldsymbol{g}(\boldsymbol{x}-(-\boldsymbol{y}))\,\psi_n(-\boldsymbol{y})\,dy = -\boldsymbol{g}_n(\boldsymbol{x})$$

Thus $\boldsymbol{g}_n$ is odd and is infinitely differentiable. Let n be large enough that

$$\|\boldsymbol{g}_n - \boldsymbol{g}\|_{\infty,\overline{\Omega}} < \delta < \operatorname{dist}(\boldsymbol{f}(\partial\Omega), \boldsymbol{0})$$

Then by definition of the degree, $d(\boldsymbol{f},\Omega,\boldsymbol{0}) = d(\boldsymbol{g},\Omega,\boldsymbol{0}) = d(\boldsymbol{g}_n,\Omega,\boldsymbol{0})$ and by Lemma 15.4.3 this is an odd integer. ■

15.5 Some Applications

With Borsuk's theorem it is possible to give relatively easy proofs of some very important and difficult theorems.

Definition 15.5.1 *If $\boldsymbol{f} : U \subseteq \mathbb{R}^p \to \mathbb{R}^p$ where U is an open set. Then $\boldsymbol{f}$ is locally one to one if for every $\boldsymbol{x} \in U$, there exists $\delta > 0$ such that $\boldsymbol{f}$ is one to one on $B(\boldsymbol{x},\delta)$.*

As a first application, consider the invariance of domain theorem. This result says that a one to one continuous map takes open sets to open sets. It is an amazing result which is essential to understand if you wish to study manifolds. In fact, the following theorem only requires $\boldsymbol{f}$ to be locally one to one. First here is a lemma which has the main idea.

Lemma 15.5.2 *Let $\boldsymbol{g} : \overline{B(\boldsymbol{0},r)} \to \mathbb{R}^p$ be one to one and continuous where here $B(\boldsymbol{0},r)$ is the ball centered at $\boldsymbol{0}$ of radius r in $\mathbb{R}^p$. Then there exists $\delta > 0$ such that*

$$\boldsymbol{g}(\boldsymbol{0}) + B(\boldsymbol{0},\delta) \subseteq \boldsymbol{g}(B(\boldsymbol{0},r)).$$

The symbol on the left means: $\{\boldsymbol{g}(\boldsymbol{0}) + \boldsymbol{x} : \boldsymbol{x} \in B(\boldsymbol{0},\delta)\}$.

Proof: For $t \in [0,1]$, let

$$\boldsymbol{h}(\boldsymbol{x},t) \equiv \boldsymbol{g}\left(\frac{\boldsymbol{x}}{1+t}\right) - \boldsymbol{g}\left(\frac{-t\boldsymbol{x}}{1+t}\right)$$

Then for $\boldsymbol{x} \in \partial B(\boldsymbol{0},r)$, $\boldsymbol{h}(\boldsymbol{x},t) \neq \boldsymbol{0}$ because if this were so, the fact $\boldsymbol{g}$ is one to one implies

$$\frac{\boldsymbol{x}}{1+t} = \frac{-t\boldsymbol{x}}{1+t}$$

and this requires $\boldsymbol{x} = \boldsymbol{0}$, not the case since $\|\boldsymbol{x}\| = r$. Then $d(\boldsymbol{h}(\cdot,t), B(\boldsymbol{0},r), \boldsymbol{0})$ is constant. Hence it is an odd integer for all t thanks to Borsuk's theorem, because $\boldsymbol{h}(\cdot,1)$ is

odd. Now let $B(\mathbf{0},\delta)$ be such that $B(\mathbf{0},\delta)\cap \boldsymbol{h}(\partial\Omega,0)=\emptyset$. Then $d(\boldsymbol{h}(\cdot,0),B(\mathbf{0},r),\mathbf{0}) = d(\boldsymbol{h}(\cdot,0),B(\mathbf{0},r),\boldsymbol{z})$ for $\boldsymbol{z}\in B(\mathbf{0},\delta)$ because the degree is constant on connected components of $\mathbb{R}^p\setminus \boldsymbol{h}(\partial\Omega,0)$. Hence $\boldsymbol{z}=\boldsymbol{h}(\boldsymbol{x},0)=\boldsymbol{g}(\boldsymbol{x})-\boldsymbol{g}(\mathbf{0})$ for some $\boldsymbol{x}\in B(0,r)$. Thus $\boldsymbol{g}(B(\mathbf{0},r))\supseteq \boldsymbol{g}(\mathbf{0})+B(\mathbf{0},\delta)$ ■

Now with this lemma, it is easy to prove the very important invariance of domain theorem.

A function $\boldsymbol{f}$ is locally one to one on an open set Ω if for every $\boldsymbol{x}_0\in\Omega$, there exists $B(\boldsymbol{x}_0,r)\subseteq\Omega$ such that $\boldsymbol{f}$ is one to one on $B(\boldsymbol{x}_0,r)$.

Theorem 15.5.3 *(invariance of domain)Let Ω be any open subset of $\mathbb{R}^p$ and let $\boldsymbol{f}:\Omega\to\mathbb{R}^p$ be continuous and locally one to one. Then $\boldsymbol{f}$ maps open subsets of Ω to open sets in $\mathbb{R}^p$.*

Proof: Let $\overline{B(\boldsymbol{x}_0,r)}\subseteq\Omega$ where $\boldsymbol{f}$ is one to one on $\overline{B(\boldsymbol{x}_0,r)}$. Let $\boldsymbol{g}$ be defined on $\overline{B(\mathbf{0},r)}$ given by

$$\boldsymbol{g}(\boldsymbol{x})\equiv \boldsymbol{f}(\boldsymbol{x}+\boldsymbol{x}_0)$$

Then $\boldsymbol{g}$ satisfies the conditions of Lemma 15.5.2, being one to one and continuous. It follows from that lemma there exists $\delta>0$ such that

$$\begin{aligned}\boldsymbol{f}(\Omega) &\supseteq \boldsymbol{f}(B(\boldsymbol{x}_0,r)) = \boldsymbol{f}(\boldsymbol{x}_0+B(\mathbf{0},r))\\ &= \boldsymbol{g}(B(\mathbf{0},r))\supseteq \boldsymbol{g}(\mathbf{0})+B(\mathbf{0},\delta)\\ &= \boldsymbol{f}(\boldsymbol{x}_0)+B(\mathbf{0},\delta)=B(\boldsymbol{f}(\boldsymbol{x}_0),\delta)\end{aligned}$$

This shows that for any $\boldsymbol{x}_0\in\Omega$, $\boldsymbol{f}(\boldsymbol{x}_0)$ is an interior point of $\boldsymbol{f}(\Omega)$ which shows $\boldsymbol{f}(\Omega)$ is open. ■

With the above, one gets easily the following amazing result. It is something which is clear for linear maps but this is a statement about continuous maps.

Corollary 15.5.4 *If $p>m$ there does not exist a continuous one to one map from $\mathbb{R}^p$ to $\mathbb{R}^m$.*

Proof: Suppose not and let $\boldsymbol{f}$ be such a continuous map,

$$\boldsymbol{f}(\boldsymbol{x})\equiv (f_1(\boldsymbol{x}),\cdots,f_m(\boldsymbol{x}))^T.$$

Then let $\boldsymbol{g}(\boldsymbol{x})\equiv (f_1(\boldsymbol{x}),\cdots,f_m(\boldsymbol{x}),0,\cdots,0)^T$ where there are $p-m$ zeros added in. Then $\boldsymbol{g}$ is a one to one continuous map from $\mathbb{R}^p$ to $\mathbb{R}^p$ and so $\boldsymbol{g}(\mathbb{R}^p)$ would have to be open from the invariance of domain theorem and this is not the case. ■

Corollary 15.5.5 *If $\boldsymbol{f}$ is locally one to one and continuous, $\boldsymbol{f}:\mathbb{R}^p\to\mathbb{R}^p$, and*

$$\lim_{|\boldsymbol{x}|\to\infty}|\boldsymbol{f}(\boldsymbol{x})|=\infty,$$

then $\boldsymbol{f}$ maps $\mathbb{R}^p$ onto $\mathbb{R}^p$.

Proof: By the invariance of domain theorem, $\boldsymbol{f}(\mathbb{R}^p)$ is an open set. It is also true that $\boldsymbol{f}(\mathbb{R}^p)$ is a closed set. Here is why. If $\boldsymbol{f}(\boldsymbol{x}_k)\to\boldsymbol{y}$, the growth condition ensures that $\{\boldsymbol{x}_k\}$ is a bounded sequence. Taking a subsequence which converges to $\boldsymbol{x}\in\mathbb{R}^p$ and using the continuity of $\boldsymbol{f}$, it follows $\boldsymbol{f}(\boldsymbol{x})=\boldsymbol{y}$. Thus $\boldsymbol{f}(\mathbb{R}^p)$ is both open and closed which implies $\boldsymbol{f}$ must be an onto map since otherwise, $\mathbb{R}^p$ would not be connected. ■

The proofs of the next two theorems make use of the Tietze extension theorem, Theorem 5.7.5.

Theorem 15.5.6 *Let Ω be a symmetric open set in $\mathbb{R}^p$ such that $\mathbf{0} \in \Omega$ and let $\boldsymbol{f} : \partial\Omega \to V$ be continuous where V is an m dimensional subspace of $\mathbb{R}^p, m < p$. Then $\boldsymbol{f}(-\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x})$ for some $\boldsymbol{x} \in \partial\Omega$.*

Proof: You could reduce to the case where $V = \mathbb{R}^m$ if desired. Suppose not. Using the Tietze extension theorem on components of the function, extend $\boldsymbol{f}$ to all of $\mathbb{R}^p$, $\boldsymbol{f}\left(\overline{\Omega}\right) \subseteq V$. (Here the extended function is also denoted by $\boldsymbol{f}$.) Let $\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(-\boldsymbol{x})$. Then $\mathbf{0} \notin \boldsymbol{g}(\partial\Omega)$ and so for some $r > 0$, $B(\mathbf{0},r) \subseteq \mathbb{R}^p \setminus \boldsymbol{g}(\partial\Omega)$. For $\boldsymbol{z} \in B(\mathbf{0},r)$,

$$d(\boldsymbol{g},\Omega,\boldsymbol{z}) = d(\boldsymbol{g},\Omega,\mathbf{0}) \neq 0$$

because $B(\mathbf{0},r)$ is contained in a component of $\mathbb{R}^p \setminus \boldsymbol{g}(\partial\Omega)$ and Borsuk's theorem implies that $d(\boldsymbol{g},\Omega,\mathbf{0}) \neq 0$ since $\boldsymbol{g}$ is odd. Hence

$$V \supseteq \boldsymbol{g}(\Omega) \supseteq B(\mathbf{0},r)$$

and this is a contradiction because V is m dimensional. ■

This theorem is called the Borsuk Ulam theorem. Note that it implies there exist two points on opposite sides of the surface of the earth which have the same atmospheric pressure and temperature, assuming the earth is symmetric and that pressure and temperature are continuous functions. The next theorem is an amusing result which is like combing hair. It gives the existence of a "cowlick".

Theorem 15.5.7 *Let p be odd and let Ω be an open bounded set in $\mathbb{R}^p$ with $\mathbf{0} \in \Omega$. Suppose $\boldsymbol{f} : \partial\Omega \to \mathbb{R}^p \setminus \{\mathbf{0}\}$ is continuous. Then for some $\boldsymbol{x} \in \partial\Omega$ and $\lambda \neq 0$, $\boldsymbol{f}(\boldsymbol{x}) = \lambda\boldsymbol{x}$.*

Proof: Using the Tietze extension theorem, extend $\boldsymbol{f}$ to all of $\mathbb{R}^p$. Also denote the extended function by $\boldsymbol{f}$. Suppose for all $\boldsymbol{x} \in \partial\Omega$, $\boldsymbol{f}(\boldsymbol{x}) \neq \lambda\boldsymbol{x}$ for all $\lambda \in \mathbb{R}$. Then

$$\mathbf{0} \notin t\boldsymbol{f}(\boldsymbol{x}) + (1-t)\boldsymbol{x}, \ \ (\boldsymbol{x},t) \in \partial\Omega \times [0,1].$$

$$\mathbf{0} \notin t\boldsymbol{f}(\boldsymbol{x}) - (1-t)\boldsymbol{x}, \ \ (\boldsymbol{x},t) \in \partial\Omega \times [0,1].$$

Thus there exists a homotopy of $\boldsymbol{f}$ and I and a homotopy of $\boldsymbol{f}$ and $-I$. Then by the homotopy invariance of degree,

$$d(\boldsymbol{f},\Omega,\mathbf{0}) = d(I,\Omega,\mathbf{0}),\ d(\boldsymbol{f},\Omega,\mathbf{0}) = d(-I,\Omega,\mathbf{0}).$$

But this is impossible because $d(I,\Omega,\mathbf{0}) = 1$ but $d(-I,\Omega,\mathbf{0}) = (-1)^n = -1$. ■

15.6 Product Formula, Separation Theorem

This section is on the product formula for the degree which is used to prove the Jordan separation theorem. To begin with is a significant observation which is used without comment below. Recall that the connected components of an open set are open. The formula is all about the composition of continuous functions.

$$\Omega \xrightarrow{\boldsymbol{f}} \boldsymbol{f}(\Omega) \subseteq \mathbb{R}^p \xrightarrow{\boldsymbol{g}} \mathbb{R}^p$$

Lemma 15.6.1 *Let $\{K_i\}_{i=1}^N, N \leq \infty$ be the connected components of $\mathbb{R}^p \setminus C$ where C is a closed set. Then $\partial K_i \subseteq C$.*

Proof: Since K_i is a connected component of an open set, it is itself open. See Theorem 3.11.12. Thus ∂K_i consists of all limit points of K_i which are not in K_i. Let $\boldsymbol{p}$ be such a point. If it is not in C then it must be in some other K_j which is impossible because these are disjoint open sets. Thus if $\boldsymbol{x}$ is a point in U it cannot be a limit point of V for V disjoint from U. ■

Definition 15.6.2 *Let the connected components of $\mathbb{R}^p \setminus \boldsymbol{f}(\partial\Omega)$ be denoted by K_i. From the properties of the degree listed in Theorem 15.2.2, $d(\boldsymbol{f},\Omega,\cdot)$ is constant on each of these components. Denote by $d(\boldsymbol{f},\Omega,K_i)$ the constant value on the component K_i.*

The following is the product formula. Note that if K is an unbounded component of $\boldsymbol{f}(\partial\Omega)^C$, then $d(\boldsymbol{f},\Omega,\boldsymbol{y}) = 0$ for all $\boldsymbol{y} \in K$ by homotopy invariance and the fact that for large enough $\|\boldsymbol{y}\|$, $\boldsymbol{f}^{-1}(\boldsymbol{y}) = \emptyset$ since $\boldsymbol{f}(\overline{\Omega})$ is compact.

Theorem 15.6.3 *(product formula)Let $\{K_i\}_{i=1}^{\infty}$ be the bounded components of $\mathbb{R}^p \setminus \boldsymbol{f}(\partial\Omega)$ for $\boldsymbol{f} \in C(\overline{\Omega};\mathbb{R}^p)$, let $\boldsymbol{g} \in C(\mathbb{R}^p,\mathbb{R}^p)$, and suppose that $\boldsymbol{y} \notin \boldsymbol{g}(\boldsymbol{f}(\partial\Omega))$ or in other words, $\boldsymbol{g}^{-1}(\boldsymbol{y}) \cap \boldsymbol{f}(\partial\Omega) = \emptyset$. Then*

$$d(\boldsymbol{g}\circ\boldsymbol{f},\Omega,\boldsymbol{y}) = \sum_{i=1}^{\infty} d(\boldsymbol{f},\Omega,K_i)\, d(\boldsymbol{g},K_i,\boldsymbol{y}). \tag{15.8}$$

All but finitely many terms in the sum are zero. If there are no bounded components of $\boldsymbol{f}(\partial\Omega)^C$, then $d(\boldsymbol{g}\circ\boldsymbol{f},\Omega,\boldsymbol{y}) = 0$.

Proof: The compact set $\boldsymbol{f}(\overline{\Omega}) \cap \boldsymbol{g}^{-1}(\boldsymbol{y})$ is contained in $\mathbb{R}^p \setminus \boldsymbol{f}(\partial\Omega)$ and so, $\boldsymbol{f}(\overline{\Omega}) \cap \boldsymbol{g}^{-1}(\boldsymbol{y})$ is covered by finitely many of the components K_j one of which may be the unbounded component. Since these components are disjoint, the other components fail to intersect $\boldsymbol{f}(\overline{\Omega}) \cap \boldsymbol{g}^{-1}(\boldsymbol{y})$. Thus, if K_i is one of these others, either it fails to intersect $\boldsymbol{g}^{-1}(\boldsymbol{y})$ or K_i fails to intersect $\boldsymbol{f}(\overline{\Omega})$. Thus either $d(\boldsymbol{f},\Omega,K_i) = 0$ because K_i fails to intersect $\boldsymbol{f}(\overline{\Omega})$ or $d(\boldsymbol{g},K_i,\boldsymbol{y}) = 0$ if K_i fails to intersect $\boldsymbol{g}^{-1}(\boldsymbol{y})$. Thus the sum is always a finite sum. I am using Theorem 15.2.2, the part which says that if $\boldsymbol{y} \notin \boldsymbol{h}(\overline{\Omega})$, then $d(\boldsymbol{h},\Omega,\boldsymbol{y}) = 0$. Note that $\partial K_i \subseteq \boldsymbol{f}(\partial\Omega)$ so $\boldsymbol{g}(\partial K_i) \subseteq \boldsymbol{g}(\boldsymbol{f}(\partial\Omega))$ and so $\boldsymbol{y} \notin \boldsymbol{g}(\partial K_i)$ because it is assumed that $\boldsymbol{y} \notin \boldsymbol{g}(\boldsymbol{f}(\partial\Omega))$.

Let $\tilde{\boldsymbol{g}}$ be in $C^{\infty}(\mathbb{R}^p,\mathbb{R}^p)$ and let $\|\tilde{\boldsymbol{g}} - \boldsymbol{g}\|_{\infty}$ be so small that for each of the finitely many K_i intersecting $\boldsymbol{f}(\overline{\Omega}) \cap \boldsymbol{g}^{-1}(\boldsymbol{y})$,

$$\begin{aligned} d(\boldsymbol{g}\circ\boldsymbol{f},\Omega,\boldsymbol{y}) &= d(\tilde{\boldsymbol{g}}\circ\boldsymbol{f},\Omega,\boldsymbol{y}) \\ d(\boldsymbol{g},K_i,\boldsymbol{y}) &= d(\tilde{\boldsymbol{g}},K_i,\boldsymbol{y}) \end{aligned} \tag{15.9}$$

Since $\partial K_i \subseteq \boldsymbol{f}(\partial\Omega)$, both conditions are obtained by letting

$$\|\boldsymbol{g} - \tilde{\boldsymbol{g}}\|_{\infty,\boldsymbol{f}(\overline{\Omega})} < \operatorname{dist}(\boldsymbol{y},\boldsymbol{g}(\boldsymbol{f}(\partial\Omega)))$$

By Lemma 15.1.5, there exists $\tilde{\boldsymbol{g}}$ such that $\boldsymbol{y}$ is a regular value of $\tilde{\boldsymbol{g}}$ in addition to 15.9 and $\tilde{\boldsymbol{g}}^{-1}(\boldsymbol{y}) \cap \boldsymbol{f}(\partial\Omega) = \emptyset$. Then $\tilde{\boldsymbol{g}}^{-1}(\boldsymbol{y})$ is contained in the union of the K_i along with the unbounded component(s) and by Lemma 15.1.5 $\tilde{\boldsymbol{g}}^{-1}(\boldsymbol{y})$ is countable. As discussed there, $\tilde{\boldsymbol{g}}^{-1}(\boldsymbol{y}) \cap K_i$ is finite if K_i is bounded. Let $\tilde{\boldsymbol{g}}^{-1}(\boldsymbol{y}) \cap K_i = \left\{\boldsymbol{x}_j^i\right\}_{j=1}^{m_i}, m_i \leq \infty$. m_i could only be ∞ on the unbounded component.

Now use Lemma 15.1.5 again to get $\tilde{\boldsymbol{f}}$ in $C^\infty\left(\overline{\Omega};\mathbb{R}^p\right)$ such that each $\boldsymbol{x}_j^i$ is a regular value of $\tilde{\boldsymbol{f}}$ on Ω and also $\left\|\tilde{\boldsymbol{f}}-\boldsymbol{f}\right\|_\infty$ is very small, so small that

$$d\left(\tilde{\boldsymbol{g}}\circ\tilde{\boldsymbol{f}},\Omega,\boldsymbol{y}\right)=d\left(\tilde{\boldsymbol{g}}\circ\boldsymbol{f},\Omega,\boldsymbol{y}\right)=d\left(\boldsymbol{g}\circ\boldsymbol{f},\Omega,\boldsymbol{y}\right)$$

and

$$d\left(\tilde{\boldsymbol{f}},\Omega,\boldsymbol{x}_j^i\right)=d\left(\boldsymbol{f},\Omega,\boldsymbol{x}_j^i\right)$$

for each i,j.

Thus, from the above,

$$\begin{aligned} d\left(\boldsymbol{g}\circ\boldsymbol{f},\Omega,\boldsymbol{y}\right) &= d\left(\tilde{\boldsymbol{g}}\circ\tilde{\boldsymbol{f}},\Omega,\boldsymbol{y}\right)\\ d\left(\tilde{\boldsymbol{f}},\Omega,\boldsymbol{x}_j^i\right) &= d\left(\boldsymbol{f},\Omega,\boldsymbol{x}_j^i\right)=d\left(\boldsymbol{f},\Omega,K_i\right)\\ d\left(\tilde{\boldsymbol{g}},K_i,\boldsymbol{y}\right) &= d\left(\boldsymbol{g},K_i,\boldsymbol{y}\right)\end{aligned}$$

Is $\boldsymbol{y}$ a regular value for $\tilde{\boldsymbol{g}}\circ\tilde{\boldsymbol{f}}$ on Ω? Suppose $\boldsymbol{z}\in\Omega$ and $\boldsymbol{y}=\tilde{\boldsymbol{g}}\circ\tilde{\boldsymbol{f}}(\boldsymbol{z})$ so $\tilde{\boldsymbol{f}}(\boldsymbol{z})\in\tilde{\boldsymbol{g}}^{-1}(\boldsymbol{y})$. Then $\tilde{\boldsymbol{f}}(\boldsymbol{z})=\boldsymbol{x}_j^i$ for some i,j and $D\tilde{\boldsymbol{f}}(\boldsymbol{z})^{-1}$ exists. Hence

$$D\left(\tilde{\boldsymbol{g}}\circ\tilde{\boldsymbol{f}}\right)(\boldsymbol{z})=D\tilde{\boldsymbol{g}}\left(\boldsymbol{x}_j^i\right)D\tilde{\boldsymbol{f}}(\boldsymbol{z}),$$

both linear transformations invertible. Thus $\boldsymbol{y}$ is a regular value of $\tilde{\boldsymbol{g}}\circ\tilde{\boldsymbol{f}}$ on Ω.

What of $\boldsymbol{x}_j^i$ in K_i where K_i is unbounded? As observed, the sum of $\operatorname{sgn}\left(\det D\tilde{\boldsymbol{f}}(\boldsymbol{z})\right)$ for $\boldsymbol{z}\in\tilde{\boldsymbol{f}}^{-1}\left(\boldsymbol{x}_j^i\right)$ is $d\left(\tilde{\boldsymbol{f}},\Omega,\boldsymbol{x}_j^i\right)$ and is 0 because the degree is constant on K_i which is unbounded.

From the definition of the degree, the left side of 15.8 $d\left(\boldsymbol{g}\circ\boldsymbol{f},\Omega,\boldsymbol{y}\right)$ equals

$$\sum\left\{\operatorname{sgn}\left(\det D\tilde{\boldsymbol{g}}\left(\tilde{\boldsymbol{f}}(\boldsymbol{z})\right)\right)\operatorname{sgn}\left(\det D\tilde{\boldsymbol{f}}(\boldsymbol{z})\right):\boldsymbol{z}\in\tilde{\boldsymbol{f}}^{-1}\left(\tilde{\boldsymbol{g}}^{-1}(\boldsymbol{y})\right)\right\}$$

The $\tilde{\boldsymbol{g}}^{-1}(\boldsymbol{y})$ are the $\boldsymbol{x}_j^i$. Thus the above is of the form

$$=\sum_i\sum_j\sum_{\boldsymbol{z}\in\tilde{\boldsymbol{f}}^{-1}\left(\boldsymbol{x}_j^i\right)}\operatorname{sgn}\left(\det\left(D\tilde{\boldsymbol{g}}\left(\boldsymbol{x}_j^i\right)\right)\right)\operatorname{sgn}\left(\det\left(D\tilde{\boldsymbol{f}}(\boldsymbol{z})\right)\right)$$

As mentioned, if $\boldsymbol{x}_j^i\in K_i$ an unbounded component, then

$$\sum_{\boldsymbol{z}\in\tilde{\boldsymbol{f}}^{-1}\left(\boldsymbol{x}_j^i\right)}\operatorname{sgn}\left(\det\left(D\tilde{\boldsymbol{g}}\left(\boldsymbol{x}_j^i\right)\right)\right)\operatorname{sgn}\left(\det\left(D\tilde{\boldsymbol{f}}(\boldsymbol{z})\right)\right)=0$$

and so, it suffices to only consider bounded components in what follows and the sum makes sense because there are finitely many $\boldsymbol{x}_j^i$ in bounded K_i. This also shows that if there are no bounded components of $\boldsymbol{f}\left(\partial\Omega\right)^C$, then $d\left(\boldsymbol{g}\circ\boldsymbol{f},\Omega,\boldsymbol{y}\right)=0$. Thus $d\left(\boldsymbol{g}\circ\boldsymbol{f},\Omega,\boldsymbol{y}\right)$ equals

$$\begin{aligned} &= \sum_i\sum_j\operatorname{sgn}\left(\det\left(D\tilde{\boldsymbol{g}}\left(\boldsymbol{x}_j^i\right)\right)\right)\sum_{\boldsymbol{z}\in\tilde{\boldsymbol{f}}^{-1}\left(\boldsymbol{x}_j^i\right)}\operatorname{sgn}\left(\det\left(D\tilde{\boldsymbol{f}}(\boldsymbol{z})\right)\right)\\ &= \sum_i d\left(\tilde{\boldsymbol{g}},K_i,\boldsymbol{y}\right)d\left(\tilde{\boldsymbol{f}},\Omega,K_i\right)\end{aligned}$$

To explain the last step,

$$\sum_{z\in\tilde{f}^{-1}\left(x_j^i\right)} \operatorname{sgn}\left(\det\left(D\tilde{f}(z)\right)\right) \equiv d\left(\tilde{f},\Omega,x_j^i\right) = d\left(\tilde{f},\Omega,K_i\right)$$

This proves the product formula because $\tilde{g}$ and $\tilde{f}$ were chosen close enough to f,g respectively that

$$\sum_i d\left(\tilde{f},\Omega,K_i\right) d(\tilde{g},K_i,y) = \sum_i d(f,\Omega,K_i)\, d(g,K_i,y) \blacksquare$$

Before the general Jordan separation theorem, I want to first consider the examples of most interest.

Recall that if a function f is continuous and one to one on a compact set K, then f is a homeomorphism of K and $f(K)$. Also recall that if U is a nonempty open set, the boundary of U, denoted as ∂U and meaning those points x with the property that for all $r>0$ $B(x,r)$ intersects both U and U^C, is $\overline{U}\setminus U$.

Proposition 15.6.4 *Let H be a compact set and let $f: H\to\mathbb{R}^p, p\geq 2$ be one to one and continuous so that H and $f(H)\equiv C$ are homeomorphic. Suppose H^C has only one connected component so H^C is connected. Then C^C also has only one component.*

Proof: Extend f, using the Tietze extension theorem to all of $\mathbb{R}^p$ and let g be an extension of f^{-1} to all of $\mathbb{R}^p$. Suppose K is a bounded component of C^C. Then $\partial K\subseteq C$ by Lemma 15.6.1. Hence $g(\partial K)\subseteq g(C)=H$. If Q is a bounded component of $g(\partial K)^C$ then if Q contains a point of H^C it follows that H^C is connected, has no points of H and hence no points of $g(\partial K)$ so $Q\supseteq H^C$ and Q is not bounded after all. Therefore, there are no bounded components of $g(\partial K)^C$. But by Theorem 15.6.3, for $\mathcal{Q}$ the set of bounded components of $g(\partial K)^C$, $d(f\circ g,K,z)=\sum_{Q\in\mathcal{Q}} d(g,K,Q)\, d(f,Q,z)=0$ because $\mathcal{Q}=\emptyset$. However, $d(f\circ g,K,z)=d(id,K,z)$ because $f\circ g=id$ on $\partial K\subseteq C$. Thus there is no bounded component of C^C so C^C has only one component just as H^C. $\blacksquare$

This says that if a compact set H fails to separate $\mathbb{R}^p$ and if f is continuous and one to one, then also $f(H)$ fails to separate $\mathbb{R}^p$.

It is obvious that the unit sphere S^{p-1} divides $\mathbb{R}^p$ into two disjoint open sets, the inside and the outside. The following shows that this also holds for any homeomorphic image of S^{p-1}.

Proposition 15.6.5 *Let B be the ball $B(0,1)$ with S^{p-1} its boundary, $p\geq 2$. Suppose $f: S^{p-1}\to C\equiv f\left(S^{p-1}\right)\subseteq\mathbb{R}^p$ is a homeomorphism. Then C^C also has exactly two components, one bounded and one unbounded.*

Proof: Let f denote the extension of f to all of $\mathbb{R}^p$ and let $g=f^{-1}$ on $f(\partial B)$ where g is also extended using the Tietze extension theorem to all of $\mathbb{R}^p$. Thus $0\notin g(f(\partial B))$ because if $x\in\partial B$, then $g(f(x))=x\neq 0$. By Theorem 15.2.2 and the product formula,

$$1=d(id,B,0)=d(g\circ f,B,0)=\sum_i d(f,B,K_i)\, d(g,K_i,0) \tag{15.10}$$

where the K_i are the bounded components of $f(\partial B)^C$. Thus from Lemma 15.6.1, $\partial K_i\subseteq f(\partial B)$ and so $g(\partial K_i)\subseteq g(f(\partial B))=\partial B$. It follows that $0\notin g(\partial K_i)\subseteq\partial B$. Pick a term in the sum which is nonzero. Let it involve K_i. Letting $y_i\in K_i$,

$$d(f,B,K_i)\, d(g,K_i,0)=d(f,B,y_i)\, d(g,K_i,B) \tag{15.11}$$

Indeed, B is connected and contains no points of $\boldsymbol{g}(\partial K_i) \subseteq \partial B$ so $\boldsymbol{z} \to d(\boldsymbol{g},K_i,\boldsymbol{z})$ is constant on B. I want to argue that B is the only bounded component of $\boldsymbol{g}(\partial K_i)^C$.

Let H be a bounded component of $\boldsymbol{g}(\partial K_i)^C$. Then H cannot contain any points of $\boldsymbol{g}(\partial K_i) \subseteq \partial B$. If H has any points of U, the unbounded component of ∂B^C, then U has no points of $\boldsymbol{g}(\partial K_i)$ and so it must be contained in H. Hence H is not bounded. Thus the only bounded components of $\boldsymbol{g}(\partial K_i)^C$ are contained in B. Since B is connected and has no points of $\boldsymbol{g}(\partial K_i) \subseteq \partial B$, B must be contained in exactly one bounded component of $\boldsymbol{g}(\partial K_i)^C$. Therefore this bounded component and B are equal and the only bounded component of $\boldsymbol{g}(\partial K_i)^C$ is B. Therefore in 15.11, the right side is just $d(\boldsymbol{f}\circ\boldsymbol{g},K_i,\boldsymbol{y}_i) = d(id,K_i,\boldsymbol{y}_i) = 1$. This is because $\boldsymbol{f}\circ\boldsymbol{g} = id$ on $\partial K_i \subseteq \boldsymbol{f}(\partial B)$ and when the two functions coincide on the boundary, they have the same degree by Theorem 15.2.2. It follows from 15.10 that, since each term in the sum is 1 or 0, there is exactly one term in the sum and hence exactly one bounded component of $\boldsymbol{f}(\partial B)^C$. ■

A repeat of the above proof yields the following corollary. Replace B with Ω.

Corollary 15.6.6 *Let $\Omega \subseteq \mathbb{R}^p$, $p \geq 2$be a bounded open connected set such that $\partial\Omega^C$ has two components, a bounded and an unbounded component. Suppose $\boldsymbol{f} : \partial\Omega \to C \equiv \boldsymbol{f}(\partial\Omega) \subseteq \mathbb{R}^p$ is a homeomorphism. Then C^C also has exactly two components, one bounded and one unbounded.*

As an application, here is a very interesting little result. It has to do with $d(\boldsymbol{f},\Omega,\boldsymbol{f}(\boldsymbol{x}))$ in the case where $\boldsymbol{f}$ is one to one and Ω is open and connected. You might imagine this should equal 1 or -1 based on one dimensional analogies. Recall a one to one map defined on an interval is either increasing or decreasing. It either preserves or reverses orientation. It is similar in n dimensions and it is a nice application of the Jordan separation theorem and the product formula.

Proposition 15.6.7 *Let Ω be an open connected bounded set in $\mathbb{R}^p, p \geq 2$ such that $\mathbb{R}^p \setminus \partial\Omega$ consists of two connected components. Let $\boldsymbol{f} \in C\left(\overline{\Omega};\mathbb{R}^p\right)$ be continuous and one to one. Then $\boldsymbol{f}(\Omega)$ is the bounded component of $\mathbb{R}^p \setminus \boldsymbol{f}(\partial\Omega)$ and for $\boldsymbol{y} \in \boldsymbol{f}(\Omega)$, $d(\boldsymbol{f},\Omega,\boldsymbol{y})$ either equals 1 or -1.*

Proof: By the Jordan separation theorem, Corollary 15.6.6, $\mathbb{R}^p \setminus \boldsymbol{f}(\partial\Omega)$ consists of two components, a bounded component B and an unbounded component U. Using the Tietze extention theorem, there exists $\boldsymbol{g}$ defined on $\mathbb{R}^p$ such that $\boldsymbol{g} = \boldsymbol{f}^{-1}$ on $\boldsymbol{f}\left(\overline{\Omega}\right)$. Thus on $\partial\Omega, \boldsymbol{g}\circ\boldsymbol{f} = \mathrm{id}$. It follows from this and the product formula that

$$\begin{aligned} 1 &= d(\mathrm{id},\Omega,\boldsymbol{g}(\boldsymbol{y})) = d(\boldsymbol{g}\circ\boldsymbol{f},\Omega,\boldsymbol{g}(\boldsymbol{y})) \\ &= d(\boldsymbol{g},B,\boldsymbol{g}(\boldsymbol{y}))\,d(\boldsymbol{f},\Omega,B) \end{aligned}$$

Therefore, $d(\boldsymbol{f},\Omega,B) \neq 0$ and so for every $\boldsymbol{z} \in B$, it follows $\boldsymbol{z} \in \boldsymbol{f}(\Omega)$. Thus $B \subseteq \boldsymbol{f}(\Omega)$. On the other hand, $\boldsymbol{f}(\Omega)$ cannot have points in both U and B because it is a connected set. Therefore $\boldsymbol{f}(\Omega) \subseteq B$ and this shows $B = \boldsymbol{f}(\Omega)$. Thus $d(\boldsymbol{f},\Omega,B) = d(\boldsymbol{f},\Omega,\boldsymbol{y})$ for each $\boldsymbol{y} \in B$ and the above formula shows this equals either 1 or -1 because the degree is an integer. ■

The one dimensional case also fits into this although it is easier to do by more elementary means. In the case where $n = 1$, the argument is essentially the same. There is one and only one bounded component for $\mathbb{R} \setminus f(\{a,b\})$. This shows how to generalize orientation.

It is just the degree. One could use this to describe an orientable manifold without any direct reference to differentiability.

In the case of $\boldsymbol{f}\left(S^{p-1}\right)$ one wants to verify that this is the is the boundary of both components, the bounded one and the unbounded one.

Theorem 15.6.8 *Let S^{p-1} be the unit sphere in $\mathbb{R}^p$, $p \geq 2$. Suppose $\gamma : S^{p-1} \to \Gamma \subseteq \mathbb{R}^p$ is one to one onto and continuous. Then $\mathbb{R}^p \setminus \Gamma$ consists of two components, a bounded component (called the inside) U_i and an unbounded component (called the outside), U_o. Also the boundary of each of these two components of $\mathbb{R}^p \setminus \Gamma$ is Γ and Γ has empty interior.*

Proof: γ^{-1} is continuous since S^{p-1} is compact and γ is one to one. By the Jordan separation theorem, $\mathbb{R}^p \setminus \Gamma = U_o \cup U_i$ where these on the right are the connected components of the set on the left, both open sets. Only U_i is bounded. Thus $\Gamma \cup U_i \cup U_o = \mathbb{R}^p$. Since both U_i, U_o are open, $\partial U \equiv \overline{U} \setminus U$ for U either U_o or U_i. If $\boldsymbol{x} \in \Gamma$, and is not a limit point of U_i, then there is $B(\boldsymbol{x}, r)$ which contains no points of U_i. Let S be those points $\boldsymbol{x}$ of Γ for which, $B(\boldsymbol{x}, r)$ contains no points of U_i for some $r > 0$. This S is open in Γ. Let $\hat{\Gamma}$ be $\Gamma \setminus S$. Then if $\hat{C} = \gamma^{-1}\left(\hat{\Gamma}\right)$, it follows that $\hat{C}$ is a closed set in $S^{p-1}$and is a proper subset of S^{p-1}. It is obvious that taking a relatively open set from S^{p-1} results in a compact set whose complement in $\mathbb{R}^p$ is an open connected set. By Proposition 15.6.4, $\mathbb{R}^p \setminus \hat{\Gamma}$ is also an open connected set. Start with $\boldsymbol{x} \in U_i$ and consider a continuous curve which goes from $\boldsymbol{x}$ to $\boldsymbol{y} \in U_o$ which is contained in $\mathbb{R}^p \setminus \hat{\Gamma}$. Thus the curve contains no points of $\hat{\Gamma}$. However, it must contain points of Γ which can only be in S. The first point of Γ intersected by this curve is a point in $\overline{U_i}$ and so this point of intersection is not in S after all because every ball containing it must contain points of U_i. Thus $S = \emptyset$ and every point of Γ is in $\overline{U_i}$. Similarly, every point of Γ is in $\overline{U_o}$. Thus $\Gamma \subseteq \overline{U_i} \setminus U_i$ and $\Gamma \subseteq \overline{U_o} \setminus U_o$. However, if $\boldsymbol{x} \in \overline{U_i} \setminus U_i$, then $\boldsymbol{x} \notin U_o$ because it is a limit point of U_i and so $\boldsymbol{x} \in \Gamma$. It is similar with U_o. Thus $\Gamma = \overline{U_i} \setminus U_i$ and $\Gamma = \overline{U_o} \setminus U_o$. This could not happen if Γ had an interior point. Such a point would be in Γ but would fail to be in either ∂U_i or ∂U_o. ■

When $p = 2$, this theorem is called the Jordan curve theorem.

What if γ maps $\bar{B}$ to $\mathbb{R}^p$ instead of γ only being defined on S^{p-1}? Obviously, one should be able to say a little more.

Corollary 15.6.9 *Let B be an open ball and let $\gamma : \bar{B} \to \mathbb{R}^p$ be one to one and continuous. Let U_i, U_o be as in the above theorem, the bounded and unbounded components of $\gamma(\partial B)^C$. Then $U_i = \gamma(B)$.*

Proof: This follows from Proposition 15.6.7.

Note how this yields the invariance of domain theorem. If $\boldsymbol{f}$ is one to one on U an open set, you could consider $\bar{B} \subseteq U$ and then $\boldsymbol{f}(B)$ is the bounded component of $\boldsymbol{f}(\partial B)^C$. You can do this for each ball contained in U. Thus $\boldsymbol{f}(U)$ is open.

15.7 General Jordan Separation Theorem

What follows is the general Jordan separation theorem.

Lemma 15.7.1 *Let Ω be a bounded open set in $\mathbb{R}^p$, $\boldsymbol{f} \in C\left(\overline{\Omega}; \mathbb{R}^p\right)$, and suppose the sequence $\{\Omega_i\}_{i=1}^{\infty}$ are disjoint open sets contained in Ω such that*

$$\boldsymbol{y} \notin \boldsymbol{f}\left(\overline{\Omega} \setminus \cup_{j=1}^{\infty} \Omega_j\right)$$

Then $d(\boldsymbol{f},\Omega,\boldsymbol{y}) = \sum_{j=1}^{\infty} d(\boldsymbol{f},\Omega_j,\boldsymbol{y})$ *where the sum has only finitely many terms equal to 0.*

Proof: By assumption, the compact set $\boldsymbol{f}^{-1}(\boldsymbol{y}) \equiv \left\{\boldsymbol{x} \in \overline{\Omega} : \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{y}\right\}$ has empty intersection with

$$\overline{\Omega} \setminus \cup_{j=1}^{\infty} \Omega_j$$

and so this compact set is covered by finitely many of the Ω_j, say $\{\Omega_1, \cdots, \Omega_{n-1}\}$ and

$$\boldsymbol{y} \notin \boldsymbol{f}\left(\cup_{j=n}^{\infty} \Omega_j\right).$$

By Theorem 15.2.2 and letting $O = \cup_{j=n}^{\infty} \Omega_j$,

$$d(\boldsymbol{f},\Omega,\boldsymbol{y}) = \sum_{j=1}^{n-1} d(\boldsymbol{f},\Omega_j,\boldsymbol{y}) + d(\boldsymbol{f},O,\boldsymbol{y}) = \sum_{j=1}^{\infty} d(\boldsymbol{f},\Omega_j,\boldsymbol{y})$$

because $d(\boldsymbol{f},O,\boldsymbol{y}) = 0$ as is $d(\boldsymbol{f},\Omega_j,\boldsymbol{y})$ for every $j \geq n$. ■

To help remember some symbols, here is a short diagram. I have a little trouble remembering what is what when I read the proof.

bounded components $\mathbb{R}^p \setminus C$	bounded components $\mathbb{R}^p \setminus \boldsymbol{f}(C)$
$\mathscr{K}$	$\mathscr{L}$

For $K \in \mathscr{K}$, K a bounded component of $\mathbb{R}^p \setminus C$,

bounded components of $\mathbb{R}^p \setminus \boldsymbol{f}(\partial K)$
$\mathscr{H}$
$\mathscr{L}_H$ sets of $\mathscr{L}$ contained in $H \in \mathscr{H}$
$\mathscr{H}_1$ those sets of $\mathscr{H}$ which intersect a set of $\mathscr{L}$

Note that the sets of $\mathscr{H}$ will tend to be larger than the sets of $\mathscr{L}$. The following lemma relating to the above definitions is interesting.

Lemma 15.7.2 *If a component* L *of* $\mathbb{R}^p \setminus \boldsymbol{f}(C)$ *intersects a component* H *of* $\mathbb{R}^p \setminus \boldsymbol{f}(\partial K)$, K *a component of* $\mathbb{R}^p \setminus C$, *then* $L \subseteq H$.

Proof: First note that by Lemma 15.6.1 for $K \in \mathscr{K}$, $\partial K \subseteq C$.

Next suppose $L \in \mathscr{L}$ and $L \cap H \neq \emptyset$ where $\mathscr{H}$ is as described above. Suppose $\boldsymbol{y} \in L \setminus H$ so L is not contained in H. Since $\boldsymbol{y} \in L$, it follows that $\boldsymbol{y} \notin \boldsymbol{f}(C)$ and so $\boldsymbol{y} \notin \boldsymbol{f}(\partial K)$ either, by Lemma 15.6.1. It follows that $\boldsymbol{y} \in \tilde{H}$ for some other $\tilde{H}$ a component of $\mathbb{R}^p \setminus \boldsymbol{f}(\partial K)$. Now $\mathbb{R}^p \setminus \boldsymbol{f}(\partial K) \supseteq \mathbb{R}^p \setminus \boldsymbol{f}(C)$ and so $\cup\mathscr{H} \supseteq L$ and so

$$L = \cup\{L \cap H : H \in \mathscr{H}\}$$

and these are disjoint open sets with two of them nonempty. Hence L would not be connected which is a contradiction. Hence if $L \cap H \neq \emptyset$, then $L \subseteq H$. ■

The following is the Jordan separation theorem. It is based on a use of the product formula and splitting up the sets of $\mathscr{L}$ into $\mathscr{L}_H$ for various $H \in \mathscr{H}$, those in $\mathscr{L}_H$ being the ones which intersect H thanks to the above lemma.

Theorem 15.7.3 *(Jordan separation theorem) Let $\boldsymbol{f}$ be a homeomorphism of C and $\boldsymbol{f}(C)$ where C is a compact set in $\mathbb{R}^p$. Then $\mathbb{R}^p \setminus C$ and $\mathbb{R}^p \setminus \boldsymbol{f}(C)$ have the same number of connected components.*

Proof: Denote by $\mathscr{K}$ the bounded components of $\mathbb{R}^p \setminus C$ and denote by $\mathscr{L}$, the bounded components of $\mathbb{R}^p \setminus \boldsymbol{f}(C)$. Also, using the Tietze extension theorem on components of a vector valued function, there exists $\bar{\boldsymbol{f}}$ an extension of $\boldsymbol{f}$ to all of $\mathbb{R}^p$ and let $\overline{\boldsymbol{f}^{-1}}$ be an extension of $\boldsymbol{f}^{-1}$ to all of $\mathbb{R}^p$. Pick $K \in \mathscr{K}$ and take $\boldsymbol{y} \in K$. Then $\partial K \subseteq C$ and so $\boldsymbol{y} \notin \overline{\boldsymbol{f}^{-1}}\left(\bar{\boldsymbol{f}}(\partial K)\right)$. Since $\overline{\boldsymbol{f}^{-1}} \circ \bar{\boldsymbol{f}}$ equals the identity I on ∂K, it follows from the properties of the degree that

$$1 = d(I, K, \boldsymbol{y}) = d\left(\overline{\boldsymbol{f}^{-1}} \circ \bar{\boldsymbol{f}}, K, \boldsymbol{y}\right).$$

Recall that if two functions agree on the boundary, then they have the same degree. Let $\mathscr{H}$ denote the set of bounded components of $\mathbb{R}^p \setminus \boldsymbol{f}(\partial K)$. Then

$$\cup\mathscr{H} = \mathbb{R}^p \setminus \boldsymbol{f}(\partial K) \supseteq \mathbb{R}^p \setminus \boldsymbol{f}(C)$$

Thus if $L \in \mathscr{L}$, then $L \subseteq \cup\mathscr{H}$ and so it must intersect some set H of $\mathscr{H}$. By the above Lemma 15.7.2, L is contained in this set of $\mathscr{H}$ so it is in $\mathscr{L}_H$.

By the product formula,

$$1 = d\left(\overline{\boldsymbol{f}^{-1}} \circ \bar{\boldsymbol{f}}, K, \boldsymbol{y}\right) = \sum_{H \in \mathscr{H}} d\left(\bar{\boldsymbol{f}}, K, H\right) d\left(\overline{\boldsymbol{f}^{-1}}, H, \boldsymbol{y}\right), \tag{15.12}$$

the sum being a finite sum. That is, there are finitely many H involved in the sum, the other terms being zero, this by the general result in the product formula.

What about those sets of $\mathscr{H}$ which contain no set of $\mathscr{L}$? These sets also have empty intersection with all sets of $\mathscr{L}$ and empty intersection with the unbounded component(s) of $\mathbb{R}^p \setminus \boldsymbol{f}(C)$ by Lemma 15.7.2. Therefore, for H one of these, $H \subseteq \boldsymbol{f}(C)$ because H has no points of $\mathbb{R}^p \setminus \boldsymbol{f}(C)$ which equals

$$\cup\mathscr{L} \cup \{\text{unbounded component(s) of } \mathbb{R}^p \setminus \boldsymbol{f}(C)\ \}.$$

Therefore,

$$d\left(\overline{\boldsymbol{f}^{-1}}, H, \boldsymbol{y}\right) = d\left(\boldsymbol{f}^{-1}, H, \boldsymbol{y}\right) = 0$$

because $\boldsymbol{y} \in K$ a bounded component of $\mathbb{R}^p \setminus C$, but for $\boldsymbol{u} \in H \subseteq \boldsymbol{f}(C)$, $\boldsymbol{f}^{-1}(\boldsymbol{u}) \in C$ so $\boldsymbol{f}^{-1}(\boldsymbol{u}) \neq \boldsymbol{y}$ implying that $d\left(\boldsymbol{f}^{-1}, H, \boldsymbol{y}\right) = 0$. Thus in 15.12, all such terms are zero. Then letting $\mathscr{H}_1$ be those sets of $\mathscr{H}$ which contain (intersect) some sets of $\mathscr{L}$, the above sum reduces to

$$1 = \sum_{H \in \mathscr{H}_1} d\left(\bar{\boldsymbol{f}}, K, H\right) d\left(\overline{\boldsymbol{f}^{-1}}, H, \boldsymbol{y}\right)$$

Note also that for $H \in \mathscr{H}_1$, $\overline{H} \setminus \cup\mathscr{L}_H = \overline{H} \setminus \cup\mathscr{L}$ and it has no points of $\mathbb{R}^p \setminus \boldsymbol{f}(C) = \cup\mathscr{L}$ so $\overline{H} \setminus \cup\mathscr{L}_H$ is contained in $\boldsymbol{f}(C)$. Thus $\boldsymbol{y} \notin \overline{\boldsymbol{f}^{-1}}\left(\overline{H} \setminus \cup\mathscr{L}_H\right) \subseteq C$ because $\boldsymbol{y} \notin C$.

It follows from Lemma 15.7.1 which comes from Theorem 15.2.2, the part about having disjoint open sets in Ω and getting a sum of degrees over these being equal to $d(\boldsymbol{f}, \Omega, \boldsymbol{y})$

that

$$\begin{aligned}\sum_{H\in\mathscr{H}_1} d\left(\bar{\boldsymbol{f}},K,H\right) d\left(\overline{\boldsymbol{f}^{-1}},H,\boldsymbol{y}\right) &= \sum_{H\in\mathscr{H}_1} d\left(\bar{\boldsymbol{f}},K,H\right) \sum_{L\in\mathscr{L}_H} d\left(\overline{\boldsymbol{f}^{-1}},L,\boldsymbol{y}\right)\\ &= \sum_{H\in\mathscr{H}_1}\sum_{L\in\mathscr{L}_H} d\left(\bar{\boldsymbol{f}},K,H\right) d\left(\overline{\boldsymbol{f}^{-1}},L,\boldsymbol{y}\right)\end{aligned}$$

where $\mathscr{L}_H$ are those sets of $\mathscr{L}$ contained in H.

Now $d\left(\bar{\boldsymbol{f}},K,H\right) = d\left(\bar{\boldsymbol{f}},K,L\right)$ where $L\in\mathscr{L}_H$. This is because L is an open connected subset of H and $\boldsymbol{z}\to d\left(\bar{\boldsymbol{f}},K,\boldsymbol{z}\right)$ is constant on H. Therefore,

$$\sum_{H\in\mathscr{H}_1}\sum_{L\in\mathscr{L}_H} d\left(\bar{\boldsymbol{f}},K,H\right) d\left(\overline{\boldsymbol{f}^{-1}},L,\boldsymbol{y}\right) = \sum_{H\in\mathscr{H}_1}\sum_{L\in\mathscr{L}_H} d\left(\bar{\boldsymbol{f}},K,L\right) d\left(\overline{\boldsymbol{f}^{-1}},L,\boldsymbol{y}\right)$$

As noted above, there are finitely many $H\in\mathscr{H}$ which are involved.

$$\mathbb{R}^p\setminus\boldsymbol{f}\left(C\right)\subseteq\mathbb{R}^p\setminus\boldsymbol{f}\left(\partial K\right)$$

and so every L must be contained in some $H\in\mathscr{H}_1$. It follows that the above reduces to

$$\sum_{L\in\mathscr{L}} d\left(\bar{\boldsymbol{f}},K,L\right) d\left(\overline{\boldsymbol{f}^{-1}},L,\boldsymbol{y}\right)$$

Where this is a finite sum because all but finitely many terms are 0.

Thus from 15.12,

$$1 = \sum_{L\in\mathscr{L}} d\left(\bar{\boldsymbol{f}},K,L\right) d\left(\overline{\boldsymbol{f}^{-1}},L,\boldsymbol{y}\right) = \sum_{L\in\mathscr{L}} d\left(\bar{\boldsymbol{f}},K,L\right) d\left(\overline{\boldsymbol{f}^{-1}},L,K\right) \tag{15.13}$$

Let $|\mathscr{K}|$ denote the number of components in $\mathscr{K}$ and similarly, $|\mathscr{L}|$ denotes the number of components in $\mathscr{L}$. Thus

$$|\mathscr{K}| = \sum_{K\in\mathscr{K}} 1 = \sum_{K\in\mathscr{K}}\sum_{L\in\mathscr{L}} d\left(\bar{\boldsymbol{f}},K,L\right) d\left(\overline{\boldsymbol{f}^{-1}},L,K\right)$$

Similarly, the argument taken another direction yields

$$|\mathscr{L}| = \sum_{L\in\mathscr{L}} 1 = \sum_{L\in\mathscr{L}}\sum_{K\in\mathscr{K}} d\left(\bar{\boldsymbol{f}},K,L\right) d\left(\overline{\boldsymbol{f}^{-1}},L,K\right)$$

If $|\mathscr{K}|<\infty$, then $\sum_{K\in\mathscr{K}} \overbrace{\sum_{L\in\mathscr{L}} d\left(\bar{\boldsymbol{f}},K,L\right) d\left(\overline{\boldsymbol{f}^{-1}},L,K\right)}^{1} < \infty$. The summation which equals 1 is a finite sum and so is the outside sum. Hence we can switch the order of summation and get

$$|\mathscr{K}| = \sum_{L\in\mathscr{L}}\sum_{K\in\mathscr{K}} d\left(\bar{\boldsymbol{f}},K,L\right) d\left(\overline{\boldsymbol{f}^{-1}},L,K\right) = |\mathscr{L}|$$

A similar argument applies if $|\mathscr{L}|<\infty$. Thus if one of these numbers $|\mathscr{K}|,|\mathscr{L}|$ is finite, so is the other and they are equal. This proves the theorem because if $p>1$ there is exactly one unbounded component to both $\mathbb{R}^p\setminus C$ and $\mathbb{R}^p\setminus\boldsymbol{f}\left(C\right)$ and if $p=1$ there are exactly two unbounded components. ■

15.8 Uniqueness of the Degree

The degree exists and has the above properties which allow one to prove amazing theorems. This is plenty to justify it. However, one might wonder whether, if we have **a** degree function with these properties, it will be the same? The answer is yes. It is likely possible to give a shorter list of desired properties however. Nevertheless, we want all these properties. First are some simple applications of the product formula which is one of the items which it is assumed satisfied by the degree.

Lemma 15.8.1 *Let $\boldsymbol{h}:\mathbb{R}^p\to\mathbb{R}^p$ be a homeomorphism. Then*

$$1=d\left(\boldsymbol{h},\Omega,\boldsymbol{h}\left(\boldsymbol{y}\right)\right)d\left(\boldsymbol{h}^{-1},\boldsymbol{h}\left(\Omega\right),\boldsymbol{y}\right)$$

whenever $\boldsymbol{y}\notin\partial\Omega$ for Ω a bounded open set. Similarly,

$$1=d\left(\boldsymbol{h}^{-1},\Omega,\boldsymbol{h}^{-1}\left(\boldsymbol{y}\right)\right)d\left(\boldsymbol{h},\boldsymbol{h}^{-1}\left(\Omega\right),\boldsymbol{y}\right).$$

Thus both terms in the factor equal either 1 or -1.

Proof: It is known that $\boldsymbol{y}\notin\partial\Omega$. Let $\mathscr{L}$ be the components of $\mathbb{R}^p\setminus\boldsymbol{h}\left(\partial\Omega\right)$. Thus $\boldsymbol{h}\left(\boldsymbol{y}\right)\notin\boldsymbol{h}\left(\partial\Omega\right)$. The product formula gives

$$1=d\left(\boldsymbol{h}^{-1}\circ\boldsymbol{h},\Omega,\boldsymbol{y}\right)=\sum_{L\in\mathscr{L}}d\left(\boldsymbol{h},\Omega,L\right)d\left(\boldsymbol{h}^{-1},L,\boldsymbol{y}\right)$$

$d\left(\boldsymbol{h}^{-1},L,\boldsymbol{y}\right)$ might be nonzero if $\boldsymbol{y}\in\boldsymbol{h}^{-1}\left(L\right)$. But, since $\boldsymbol{h}^{-1}$ is one to one, if $\boldsymbol{y}\in\boldsymbol{h}^{-1}\left(L\right)$, then $\boldsymbol{y}\notin\boldsymbol{h}^{-1}\left(\tilde{L}\right)$ for any other $\tilde{L}\in\mathscr{L}$. Thus there is only one term in the above sum.

$$1=d\left(\boldsymbol{h},\Omega,L\right)d\left(\boldsymbol{h}^{-1},L,\boldsymbol{y}\right),\boldsymbol{y}\in\boldsymbol{h}^{-1}\left(L\right)$$

$$1=d\left(\boldsymbol{h},\Omega,\boldsymbol{h}\left(\boldsymbol{y}\right)\right)d\left(\boldsymbol{h}^{-1},\boldsymbol{h}\left(\Omega\right),\boldsymbol{y}\right)$$

Since the degree is an integer, both factors are either 1 or -1. ■

The following assumption is convenient and seems like something which should be true. It says just a little more than $d\left(\mathrm{id},\Omega,\boldsymbol{y}\right)=1$ for $\boldsymbol{y}\in\Omega$. Actually, you could prove it by adding in a succession of $\frac{\boldsymbol{z}}{n}$ to id for n large till you get to $\boldsymbol{h}$ in the first argument and $\boldsymbol{h}_{\boldsymbol{z}}\left(\boldsymbol{y}\right)$ in the last and use the property that the degree is continuous in the first and last places. However, this is such a reasonable thing to assume, that it seems like we might as well have assumed it instead of $d\left(\mathrm{id},\Omega,\boldsymbol{y}\right)=1$ for $\boldsymbol{y}\in\Omega$.

Assumption 15.8.2 *If $\boldsymbol{h}_{\boldsymbol{z}}\left(\boldsymbol{x}\right)=\boldsymbol{x}+\boldsymbol{z}$, then for $\boldsymbol{y}\notin\partial\Omega,\boldsymbol{y}\in\Omega$,*

$$d\left(\boldsymbol{h}_{\boldsymbol{z}},\Omega,\boldsymbol{h}_{\boldsymbol{z}}\left(\boldsymbol{y}\right)\right)=d\left(\mathrm{id},\Omega,\boldsymbol{y}\right)=1.$$

That is,$d\left(\cdot+\boldsymbol{z},\Omega,\boldsymbol{z}+\boldsymbol{y}\right)=1$. From the above lemma, $d\left(\boldsymbol{h}_{\boldsymbol{z}},\boldsymbol{h}_{-\boldsymbol{z}}\left(\Omega\right),\boldsymbol{y}\right)=1$ also.

Now let $\boldsymbol{h}_{\boldsymbol{z}}\left(\boldsymbol{x}\right)\equiv\boldsymbol{x}+\boldsymbol{z}$. Let $\boldsymbol{y}\notin\boldsymbol{g}\left(\partial\Omega\right)$ and consider

$$d\left(\boldsymbol{h}_{-\boldsymbol{y}}\circ\boldsymbol{g}\circ\boldsymbol{h}_{\boldsymbol{z}},\boldsymbol{h}_{-\boldsymbol{z}}\left(\Omega\right),\boldsymbol{h}_{-\boldsymbol{y}}\left(\boldsymbol{y}\right)\right)=d\left(\boldsymbol{h}_{-\boldsymbol{y}}\circ\boldsymbol{g}\circ\boldsymbol{h}_{\boldsymbol{z}},\boldsymbol{h}_{-\boldsymbol{z}}\left(\Omega\right),\boldsymbol{0}\right)$$

Let $\mathscr{L}$ be the components of $\mathbb{R}^p\setminus\boldsymbol{g}\circ\boldsymbol{h}_{\boldsymbol{z}}\left(\partial\left(\boldsymbol{h}_{-\boldsymbol{z}}\left(\Omega\right)\right)\right)=\mathbb{R}^p\setminus\boldsymbol{g}\left(\partial\Omega\right)$. Then the product rule gives $d\left(\boldsymbol{h}_{-\boldsymbol{y}}\circ\boldsymbol{g}\circ\boldsymbol{h}_{\boldsymbol{z}},\boldsymbol{h}_{-\boldsymbol{z}}\left(\Omega\right),\boldsymbol{0}\right)=$

$$\sum_{L\in\mathscr{L}}d\left(\boldsymbol{g}\circ\boldsymbol{h}_{\boldsymbol{z}},\boldsymbol{h}_{-\boldsymbol{z}}\left(\Omega\right),L\right)d\left(\boldsymbol{h}_{-y},L,\boldsymbol{0}\right)$$

Now $\boldsymbol{h}_{-\boldsymbol{y}}$ is one to one and so if $\mathbf{0} \in \boldsymbol{h}_{-\boldsymbol{y}}(L)$, then this is true for only that single L and so there is only one term in the sum. For a single L where $\mathbf{0} \in \boldsymbol{h}_{-\boldsymbol{y}}(L)$ so $\boldsymbol{y} \in L$,

$$d\left(\boldsymbol{h}_{-\boldsymbol{y}} \circ \boldsymbol{g} \circ \boldsymbol{h}_{\boldsymbol{z}}, \boldsymbol{h}_{-\boldsymbol{z}}(\Omega), \boldsymbol{h}_{-\boldsymbol{y}}(\boldsymbol{y})\right) = d\left(\boldsymbol{g} \circ \boldsymbol{h}_{\boldsymbol{z}}, \boldsymbol{h}_{-\boldsymbol{z}}(\Omega), L\right) d\left(\boldsymbol{h}_{-y}, L, \boldsymbol{h}_{-\boldsymbol{y}}(\boldsymbol{y})\right)$$

Now from the assumption, this equals $d\left(\boldsymbol{g} \circ \boldsymbol{h}_{\boldsymbol{z}}, \boldsymbol{h}_{-\boldsymbol{z}}(\Omega), L\right) = d\left(\boldsymbol{g} \circ \boldsymbol{h}_{\boldsymbol{z}}, \boldsymbol{h}_{-\boldsymbol{z}}(\Omega), \boldsymbol{y}\right)$. Now we use the product rule on this. Letting $\mathscr{K}$ be the bounded components of

$$\mathbb{R}^p \setminus \boldsymbol{h}_{\boldsymbol{z}}\left(\partial \boldsymbol{h}_{-\boldsymbol{z}}(\Omega)\right) = \mathbb{R}^p \setminus \partial\Omega$$

$$d\left(\boldsymbol{g} \circ \boldsymbol{h}_{\boldsymbol{z}}, \boldsymbol{h}_{-\boldsymbol{z}}(\Omega), \boldsymbol{y}\right) = \sum_{K \in \mathscr{K}} d\left(\boldsymbol{h}_{\boldsymbol{z}}, \boldsymbol{h}_{-\boldsymbol{z}}(\Omega), K\right) d\left(\boldsymbol{g}, K, \boldsymbol{y}\right)$$

The points of K are not in $\partial\Omega$ and so, from the assumption, $d\left(\boldsymbol{h}_{\boldsymbol{z}}, \boldsymbol{h}_{-\boldsymbol{z}}(\Omega), K\right) = 1$. Thus this whole thing reduces to $\sum_{K \in \mathscr{K}} d\left(\boldsymbol{g}, K, \boldsymbol{y}\right) = d\left(\boldsymbol{g}, \Omega, \boldsymbol{y}\right)$. This shows the following lemma.

Lemma 15.8.3 *The following formula holds*

$$d\left(\boldsymbol{h}_{-\boldsymbol{y}} \circ \boldsymbol{g} \circ \boldsymbol{h}_{\boldsymbol{z}}, \boldsymbol{h}_{-\boldsymbol{z}}(\Omega), \boldsymbol{h}_{-\boldsymbol{y}}(\boldsymbol{y})\right) = d\left(\boldsymbol{g} \circ \boldsymbol{h}_{\boldsymbol{z}}, \boldsymbol{h}_{-\boldsymbol{z}}(\Omega), L\right) = d\left(\boldsymbol{g}, \Omega, \boldsymbol{y}\right).$$

In other words, $d\left(\boldsymbol{g}, \Omega, \boldsymbol{y}\right) = d\left(\boldsymbol{g}(\cdot + \boldsymbol{z}), \Omega - \boldsymbol{z}, \boldsymbol{y}\right) = d\left(\boldsymbol{g}(\cdot + \boldsymbol{z}) - \boldsymbol{y}, \Omega - \boldsymbol{z}, \mathbf{0}\right)$.

Theorem 15.8.4 *You have a function d which has integer values*

$$d\left(\boldsymbol{g}, \Omega, \boldsymbol{y}\right) \in \mathbb{Z}$$

whenever $\boldsymbol{y} \notin \boldsymbol{g}(\partial\Omega)$ for $\boldsymbol{g} \in C\left(\overline{\Omega}; \mathbb{R}^p\right)$. Assume it satisfies the following properties which the one above satisfies.

1. *(homotopy invariance) If*
$$\boldsymbol{h} \in C\left(\overline{\Omega} \times [0,1], \mathbb{R}^p\right)$$
and $\boldsymbol{y}(t) \notin \boldsymbol{h}(\partial\Omega, t)$ for all $t \in [0,1]$ where $\boldsymbol{y}$ is continuous, then
$$t \to d\left(\boldsymbol{h}(\cdot, t), \Omega, \boldsymbol{y}(t)\right)$$
is constant for $t \in [0,1]$.

2. *Let $\Omega \supseteq \Omega_1 \cup \Omega_2$ where $\Omega_1 \cap \Omega_2 = \emptyset$, for Ω_i an open set. Then $\boldsymbol{y} \notin \boldsymbol{f}\left(\overline{\Omega} \setminus (\Omega_1 \cup \Omega_2)\right)$, then*
$$d\left(\boldsymbol{f}, \Omega_1, \boldsymbol{y}\right) + d\left(\boldsymbol{f}, \Omega_2, \boldsymbol{y}\right) = d\left(\boldsymbol{f}, \Omega, \boldsymbol{y}\right)$$

3. *$d(\mathrm{id}, \Omega, \boldsymbol{y}) = 1$ if $\boldsymbol{y} \in \Omega$.*

4. *$d\left(\boldsymbol{f}, \Omega, \cdot\right)$ is continuous and constant on every connected component of $\mathbb{R}^p \setminus \boldsymbol{f}(\partial\Omega)$.*

5. *If $\boldsymbol{y} \notin \boldsymbol{f}(\partial\Omega)$, and if $d\left(\boldsymbol{f}, \Omega, \boldsymbol{y}\right) \neq 0$, then there exists $\boldsymbol{x} \in \Omega$ such that $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{y}$.*

6. *Product formula, Assumption 15.8.2.*

Then d is the degree which was defined above. Thus, in a sense, the degree is unique if we want it to do these things.

Proof: First note that $\boldsymbol{h} \to d(\boldsymbol{h},\Omega,\boldsymbol{y})$ is continuous on $C\left(\overline{\Omega},\mathbb{R}^p\right)$. Say

$$\|\boldsymbol{g}-\boldsymbol{h}\|_\infty < \operatorname{dist}(\boldsymbol{h}(\partial\Omega),\boldsymbol{y}).$$

Then if $\boldsymbol{x} \in \partial\Omega, t \in [0,1]$,

$$\begin{aligned}\|\boldsymbol{g}(\boldsymbol{x})+t(\boldsymbol{h}-\boldsymbol{g})(\boldsymbol{x})-\boldsymbol{y}\| &= \|\boldsymbol{g}(\boldsymbol{x})-\boldsymbol{h}(\boldsymbol{x})+t(\boldsymbol{h}-\boldsymbol{g})(\boldsymbol{x})+\boldsymbol{h}(\boldsymbol{x})-\boldsymbol{y}\| \\ &\geq \operatorname{dist}(\boldsymbol{h}(\partial\Omega),\boldsymbol{y})-(1-t)\|\boldsymbol{g}-\boldsymbol{h}\|_\infty > 0\end{aligned}$$

By homotopy invariance, $d(\boldsymbol{g},\Omega,\boldsymbol{y}) = d(\boldsymbol{h},\Omega,\boldsymbol{y})$. By the approximation lemma, if we can identify the degree for $\boldsymbol{g} \in C^\infty\left(\overline{\Omega},\mathbb{R}^p\right)$ with $\boldsymbol{y}$ a regular value, $\boldsymbol{y} \notin \boldsymbol{g}(\partial\Omega)$ then we know what the degree is. Say $\boldsymbol{g}^{-1}(\boldsymbol{y}) = \{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\}$. Using the inverse function theorem there are balls B_{ir} containing $\boldsymbol{x}_i$ such that none of these balls of radius r intersect and $\boldsymbol{g}$ is one to one on each. Then from the theorem on the fundamental properties assumed above,

$$d(\boldsymbol{g},\Omega,\boldsymbol{y}) = \sum_{i=1}^n d(\boldsymbol{g},B_{ir},\boldsymbol{y}) \tag{15.14}$$

and by assumption, this is

$$\begin{aligned}\sum_{i=1}^n d(\boldsymbol{g}(\cdot+\boldsymbol{x}_i),B(\boldsymbol{0},r),\boldsymbol{y}) &= \sum_{i=1}^n d(\boldsymbol{g}(\cdot+\boldsymbol{x}_i)-\boldsymbol{y},B(\boldsymbol{0},r),\boldsymbol{0}) \\ &= \sum_{i=1}^n d(\boldsymbol{h},B(\boldsymbol{0},r),\boldsymbol{0})\end{aligned}$$

where $\boldsymbol{h}(\boldsymbol{x}) \equiv \boldsymbol{g}(\boldsymbol{x}+\boldsymbol{x}_i)-\boldsymbol{y}$. There is no restriction on the size of r. Consider one of the terms in the sum. $d(\boldsymbol{h},B(\boldsymbol{0},r),\boldsymbol{y})$. Note $D\boldsymbol{h}(\boldsymbol{0}) = D\boldsymbol{g}(\boldsymbol{x}_i)$.

$$\boldsymbol{h}(\boldsymbol{x}) = D\boldsymbol{h}(\boldsymbol{0})\boldsymbol{x}+\boldsymbol{o}(\boldsymbol{x}) \tag{15.15}$$

Now let

$$L(\boldsymbol{z}_1,...,\boldsymbol{z}_p) \equiv \begin{pmatrix} \nabla h_1(\boldsymbol{z}_1) \\ \vdots \\ \nabla h_p(\boldsymbol{z}_p) \end{pmatrix}$$

By the mean value theorem applied to the components of $\boldsymbol{h}$, if $\boldsymbol{x} \neq \boldsymbol{0}, \|\boldsymbol{x}\| \leq r$, there exist $\boldsymbol{z}_1,...,\boldsymbol{z}_p \in \overline{B(\boldsymbol{0},r)} \times \cdots \times \overline{B(\boldsymbol{0},r)}$

$$\frac{\|\boldsymbol{h}(\boldsymbol{x})-\boldsymbol{h}(\boldsymbol{0})\|}{\|\boldsymbol{x}\|} = \frac{\|\boldsymbol{h}(\boldsymbol{x})-\boldsymbol{0}\|}{\|\boldsymbol{x}\|} =$$

$$\frac{\left\|L(\boldsymbol{z}_1,...,\boldsymbol{z}_p)\boldsymbol{x}\right\|}{\|\boldsymbol{x}\|} = \left\|L(\boldsymbol{z}_1,...,\boldsymbol{z}_n)\left(\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}\right)\right\| \tag{15.16}$$

Since $D\boldsymbol{h}(\boldsymbol{0})^{-1}$ exists, we can choose r small enough that for all $(\boldsymbol{z}_1,...,\boldsymbol{z}_p) \in \overline{B(\boldsymbol{0},r)} \times \cdots \times \overline{B(\boldsymbol{0},r)}$

$$\det(L(\boldsymbol{z}_1,...,\boldsymbol{z}_p)) \neq 0$$

Hence, for sufficiently small r, there is $\delta > 0$ such that for S the unit sphere, $\{\boldsymbol{x} : \|\boldsymbol{x}\| = 1\}$,

$$\inf\left\{ \begin{array}{c} \left\|L(\boldsymbol{z}_1,...,\boldsymbol{z}_p)\boldsymbol{x}\right\| : (\boldsymbol{z}_1,...,\boldsymbol{z}_p,\boldsymbol{x}) \\ \in \overline{B(\boldsymbol{0},r)} \times \cdots \times \overline{B(\boldsymbol{0},r)} \times S \end{array} \right\} = \delta > 0$$

Then it follows from 15.16 that if $\|\boldsymbol{x}\| \leq r$, $\frac{\|\boldsymbol{h}(\boldsymbol{x})\|}{\|\boldsymbol{x}\|} \geq \delta$ so $\|\boldsymbol{h}(\boldsymbol{x}) - \mathbf{0}\| \geq \delta \|\boldsymbol{x}\|$. Letting $\|\boldsymbol{x}\| = r$ so $\boldsymbol{x} \in \boldsymbol{h}(\partial B(\mathbf{0},r))$,

$$\operatorname{dist}(\boldsymbol{h}(\partial B(\mathbf{0},r)), \mathbf{0}) \geq \delta r$$

Now pick $\varepsilon < \delta$. Making r still smaller if necessary, it follows from 15.15 that if $\|\boldsymbol{x}\| \leq r$,

$$\|\boldsymbol{h}(\boldsymbol{x}) - D\boldsymbol{h}(\mathbf{0})\boldsymbol{x}\| \leq \varepsilon \|\boldsymbol{x}\| < \delta r \leq \operatorname{dist}(\boldsymbol{h}(\partial B(\mathbf{0},r)), \mathbf{0})$$

Thus,

$$\|\boldsymbol{h} - D\boldsymbol{h}(\mathbf{0})(\cdot)\|_{\infty, \overline{B(\mathbf{0},r)}} < \operatorname{dist}(\boldsymbol{h}(\partial B(\mathbf{0},r)), \mathbf{0})$$

and so $d(\boldsymbol{h}, B(\mathbf{0},r), \mathbf{0}) = d(D\boldsymbol{h}(\mathbf{0})(\cdot), B(\mathbf{0},r), \mathbf{0})$.

Say $A = D\boldsymbol{h}(\mathbf{0})$ an invertible matrix. Then if U is any open set containing $\mathbf{0}$, it follows from the properties of the degree that $d(A, U, \mathbf{0}) = d(A, B(\mathbf{0},r), \mathbf{0})$ whenever r is small enough. By the product formula and 3. above,

$$1 = d(I, B(\mathbf{0},r), \mathbf{0}) = d\left(A^{-1}A, B(\mathbf{0},r), \mathbf{0}\right)$$

$$\begin{aligned} = \quad & d(A, B(\mathbf{0},r), AB(\mathbf{0},r))\, d\left(A^{-1}, AB(\mathbf{0},r), \mathbf{0}\right) \\ & + d\left(A, B(\mathbf{0},r), \overline{AB(\mathbf{0},r)}^{C}\right) d\left(A^{-1}, \overline{AB(\mathbf{0},r)}^{C}, \mathbf{0}\right) \end{aligned}$$

From 5. above, this reduces to

$$1 = d(A, B(\mathbf{0},r), AB(\mathbf{0},r))\, d\left(A^{-1}, AB(\mathbf{0},r), \mathbf{0}\right)$$

Thus, since the degree is an integer, either both

$$d(A, B(\mathbf{0},r), \mathbf{0}), d\left(A^{-1}, AB(\mathbf{0},r), \mathbf{0}\right)$$

are 1 or they are both -1. This would hold if when A is invertible,

$$d(A, B(\mathbf{0},r), \mathbf{0}) = \operatorname{sgn}(\det(A)) = \operatorname{sgn}\left(\det\left(A^{-1}\right)\right).$$

It would also hold if $d(A, B(\mathbf{0},r), \mathbf{0}) = -\operatorname{sgn}(\det(A))$. However, the latter of the two alternatives is not the one wanted because, doing something like the above,

$$\begin{aligned} d\left(A^2, B(\mathbf{0},r), \mathbf{0}\right) &= d(A, B(\mathbf{0},r), AB(\mathbf{0},r))\, d(A, AB(\mathbf{0},r), \mathbf{0}) \\ &= d(A, B(\mathbf{0},r), \mathbf{0})\, d(A, AB(\mathbf{0},r), \mathbf{0}) \end{aligned}$$

If you had for a definition $d(A, B(\mathbf{0},r), \mathbf{0}) = -\operatorname{sgn}\det(A)$, then you would have

$$-\operatorname{sgn}\det\left(A^2\right) = -1 = (-\operatorname{sgn}\det(A))^2 = 1$$

Hence the only reasonable definition is to let $d(A, B(\mathbf{0},r), \mathbf{0}) = \operatorname{sgn}(\det(A))$. It follows from 15.14 that $d(\boldsymbol{g}, \Omega, \boldsymbol{y}) = \sum_{i=1}^{n} d(\boldsymbol{g}, B_{ir}, \boldsymbol{y}) = \sum_{i=1}^{n} \operatorname{sgn}(\det(D\boldsymbol{g}(\boldsymbol{x}_i)))$. Thus, if the conditions given above hold, then in whatever manner you construct the degree, it amounts to the definition given above in this chapter in the sense that for $\boldsymbol{y}$ a regular point of a smooth function, you get the definition of the chapter. ■

15.9 Exercises

1. Show that if $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_r$ in $\mathbb{R}^p \setminus \boldsymbol{f}(\partial\Omega)$, then if $\tilde{\boldsymbol{f}}$ has the property that

$$\left\| \tilde{\boldsymbol{f}} - \boldsymbol{f} \right\|_\infty < \min_{i \leq r} \operatorname{dist}(\boldsymbol{y}_i, \boldsymbol{f}(\partial\Omega)),$$

then $d(\boldsymbol{f}, \Omega, \boldsymbol{y}_i) = d\left(\tilde{\boldsymbol{f}}, \Omega, \boldsymbol{y}_i\right)$ for each $\boldsymbol{y}_i$. **Hint:** Consider for

$$t \in [0,1], \boldsymbol{f}(\boldsymbol{x}) + t\left(\tilde{\boldsymbol{f}}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x})\right) - \boldsymbol{y}_i$$

and homotopy invariance.

2. Show the Brouwer fixed point theorem is equivalent to the nonexistence of a continuous retraction onto the boundary of $B(\mathbf{0}, r)$.

3. Give a version of Proposition 15.6.7 which is valid for the case where $n = 1$.

4. It was shown that if $\lim_{|\boldsymbol{x}| \to \infty} |\boldsymbol{f}(\boldsymbol{x})| = \infty$, $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^n$ is locally one to one and continuous, then $\boldsymbol{f}$ maps $\mathbb{R}^n$ onto $\mathbb{R}^n$. Suppose you have $\boldsymbol{f} : \mathbb{R}^m \to \mathbb{R}^n$ where $\boldsymbol{f}$ is one to one, continuous, and $\lim_{|\boldsymbol{x}| \to \infty} |\boldsymbol{f}(\boldsymbol{x})| = \infty$, $m < n$. Show that $\boldsymbol{f}$ cannot be onto.

5. Can there exist a one to one onto continuous map, $\boldsymbol{f}$ which takes the unit interval to the unit disk?

6. Let $m < n$ and let $B_m(\mathbf{0}, r)$ be the ball in $\mathbb{R}^m$ and $B_n(\mathbf{0}, r)$ be the ball in $\mathbb{R}^n$. Show that there is no one to one continuous map from $\overline{B_m(\mathbf{0}, r)}$ to $\overline{B_n(\mathbf{0}, r)}$. **Hint:** It is like the above problem.

7. Consider the unit disk, $\left\{(x,y) : x^2 + y^2 \leq 1\right\} \equiv D$ and the annulus

$$\left\{(x,y) : \frac{1}{2} \leq x^2 + y^2 \leq 1\right\} \equiv A.$$

Is it possible there exists a one to one onto continuous map $\boldsymbol{f}$ such that $\boldsymbol{f}(D) = A$? Thus D has no holes and A is really like D but with one hole punched out. Can you generalize to different numbers of holes? **Hint:** Consider the invariance of domain theorem. The interior of D would need to be mapped to the interior of A. Where do the points of the boundary of A come from? Consider Theorem 3.11.3.

8. Suppose C is a compact set in $\mathbb{R}^n$ which has empty interior and $\boldsymbol{f} : C \to \Gamma \subseteq \mathbb{R}^n$ is one to one onto and continuous with continuous inverse. Could Γ have nonempty interior? Show also that if $\boldsymbol{f}$ is one to one and onto Γ then if it is continuous, so is $\boldsymbol{f}^{-1}$.

9. Let K be a nonempty closed and convex subset of $\mathbb{R}^p$. Recall K is convex means that if $\boldsymbol{x}, \boldsymbol{y} \in K$, then for all $t \in [0,1]$, $t\boldsymbol{x} + (1-t)\boldsymbol{y} \in K$. Show that if $\boldsymbol{x} \in \mathbb{R}^p$ there exists a unique $\boldsymbol{z} \in K$ such that $|\boldsymbol{x} - \boldsymbol{z}| = \min\{|\boldsymbol{x} - \boldsymbol{y}| : \boldsymbol{y} \in K\}$. This $\boldsymbol{z}$ will be denoted as $P\boldsymbol{x}$. **Hint:** First note you do not know K is compact. Establish the parallelogram identity if you have not already done so,

$$|\boldsymbol{u} - \boldsymbol{v}|^2 + |\boldsymbol{u} + \boldsymbol{v}|^2 = 2|\boldsymbol{u}|^2 + 2|\boldsymbol{v}|^2.$$

Then let $\{z_k\}$ be a minimizing sequence,

$$\lim_{k\to\infty} |z_k - x|^2 = \inf\{|x-y| : y \in K\} \equiv \lambda.$$

Using convexity, explain why

$$\left|\frac{z_k - z_m}{2}\right|^2 + \left|x - \frac{z_k + z_m}{2}\right|^2 = 2\left|\frac{x - z_k}{2}\right|^2 + 2\left|\frac{x - z_m}{2}\right|^2$$

and then use this to argue $\{z_k\}$ is a Cauchy sequence. Then if z_i works for $i = 1,2$, consider $(z_1 + z_2)/2$ to get a contradiction.

10. In Problem 9 show that Px satisfies the following variational inequality. $(x - Px) \cdot (y - Px) \leq 0$ for all $y \in K$. Then show that $|Px_1 - Px_2| \leq |x_1 - x_2|$. **Hint:** For the first part note that if $y \in K$, the function

$$t \to |x - (Px + t(y - Px))|^2$$

achieves its minimum on $[0,1]$ at $t = 0$. For the second part,

$$(x_1 - Px_1) \cdot (Px_2 - Px_1) \leq 0,\ (x_2 - Px_2) \cdot (Px_1 - Px_2) \leq 0.$$

Explain why

$$(x_2 - Px_2 - (x_1 - Px_1)) \cdot (Px_2 - Px_1) \geq 0$$

and then use a some manipulations and the Cauchy Schwarz inequality to get the desired inequality.

11. Suppose D is a set which is homeomorphic to $\overline{B(0,1)}$. This means there exists a continuous one to one map, h such that $h\left(\overline{B(0,1)}\right) = D$ such that h^{-1} is also one to one. Show that if f is a continuous function which maps D to D then f has a fixed point. Now show that it suffices to say that h is one to one and continuous. In this case the continuity of h^{-1} is automatic. Sets which have the property that continuous functions taking the set to itself have at least one fixed point are said to have the fixed point property. Work Problem 7 using this notion of fixed point property. What about a solid ball and a donut? Could these be homeomorphic?

12. Using the definition of the derivative and the Vitali covering theorem, show that if $f \in C^1\left(\overline{U}, \mathbb{R}^n\right)$ and ∂U has n dimensional measure zero then $f(\partial U)$ also has measure zero. (This problem has little to do with this chapter. It is a review.)

13. Suppose Ω is any open bounded subset of $\mathbb{R}^n$ which contains $\mathbf{0}$ and that $f : \overline{\Omega} \to \mathbb{R}^n$ is continuous with the property that $f(x) \cdot x \geq 0$ for all $x \in \partial\Omega$. Show that then there exists $x \in \Omega$ such that $f(x) = \mathbf{0}$. Give a similar result in the case where the above inequality is replaced with $\leq$. **Hint:** You might consider the function $h(t,x) \equiv t f(x) + (1-t)x$.

14. Suppose Ω is an open set in $\mathbb{R}^n$ containing $\mathbf{0}$ and suppose that $f : \overline{\Omega} \to \mathbb{R}^n$ is continuous and $|f(x)| \leq |x|$ for all $x \in \partial\Omega$. Show f has a fixed point in $\overline{\Omega}$. **Hint:** Consider $h(t,x) \equiv t(x - f(x)) + (1-t)x$ for $t \in [0,1]$. If $t = 1$ and some $x \in \partial\Omega$ is sent to $\mathbf{0}$, then you are done. Suppose therefore, that no fixed point exists on $\partial\Omega$. Consider $t < 1$ and use the given inequality.

15. Let Ω be an open bounded subset of $\mathbb{R}^n$ and let $\boldsymbol{f},\boldsymbol{g}:\overline{\Omega}\to\mathbb{R}^n$ both be continuous, $\mathbf{0}\notin\boldsymbol{f}(\partial\Omega)$, such that $|\boldsymbol{f}(\boldsymbol{x})|-|\boldsymbol{g}(\boldsymbol{x})|>0$ for all $\boldsymbol{x}\in\partial\Omega$. Show that then $d(\boldsymbol{f}-\boldsymbol{g},\Omega,\mathbf{0})=d(\boldsymbol{f},\Omega,\mathbf{0})$. Show that if there exists $\boldsymbol{x}\in\boldsymbol{f}^{-1}(\mathbf{0})$, then there exists $\boldsymbol{x}\in(\boldsymbol{f}-\boldsymbol{g})^{-1}(\mathbf{0})$. **Hint:** Consider $\boldsymbol{h}(t,\boldsymbol{x})\equiv(1-t)\boldsymbol{f}(\boldsymbol{x})+t(\boldsymbol{f}(\boldsymbol{x})-\boldsymbol{g}(\boldsymbol{x}))$ and argue $\mathbf{0}\notin\boldsymbol{h}(t,\partial\Omega)$ for $t\in[0,1]$.

16. Let $f:\mathbb{C}\to\mathbb{C}$ where $\mathbb{C}$ is the field of complex numbers. Thus f has a real and imaginary part. Letting $z=x+iy$, $f(z)=u(x,y)+iv(x,y)$. Recall that the norm in $\mathbb{C}$ is given by $|x+iy|=\sqrt{x^2+y^2}$ and this is the usual norm in $\mathbb{R}^2$ for the ordered pair (x,y). Thus complex valued functions defined on $\mathbb{C}$ can be considered as $\mathbb{R}^2$ valued functions defined on some subset of $\mathbb{R}^2$. Such a complex function is said to be analytic if the usual definition holds. That is $f'(z)=\lim_{h\to0}\frac{f(z+h)-f(z)}{h}$. In other words,
$$f(z+h)=f(z)+f'(z)h+o(h) \tag{15.17}$$
at a point z where the derivative exists. Let $f(z)=z^n$ where n is a positive integer. Thus $z^n=p(x,y)+iq(x,y)$ for p,q suitable polynomials in x and y. Show this function is analytic. Next show that for an analytic function and u and v the real and imaginary parts, the Cauchy Riemann equations hold, $u_x=v_y$, $u_y=-v_x$. In terms of mappings show 15.17 has the form
$$\begin{pmatrix} u(x+h_1,y+h_2)\\ v(x+h_1,y+h_2)\end{pmatrix}$$
$$\begin{aligned}
&= \begin{pmatrix} u(x,y)\\ v(x,y)\end{pmatrix}+\begin{pmatrix} u_x(x,y) & u_y(x,y)\\ v_x(x,y) & v_y(x,y)\end{pmatrix}\begin{pmatrix} h_1\\ h_2\end{pmatrix}+\boldsymbol{o}(\boldsymbol{h})\\
&= \begin{pmatrix} u(x,y)\\ v(x,y)\end{pmatrix}+\begin{pmatrix} u_x(x,y) & -v_x(x,y)\\ v_x(x,y) & u_x(x,y)\end{pmatrix}\begin{pmatrix} h_1\\ h_2\end{pmatrix}+\boldsymbol{o}(\boldsymbol{h})
\end{aligned}$$
where $\boldsymbol{h}=(h_1,h_2)^T$ and h is given by h_1+ih_2. Thus the determinant of the above matrix is always nonnegative. Letting B_r denote the ball $B(0,r)=B((0,0),r)$ show $d(f,B_r,\mathbf{0})=n$ where $f(z)=z^n$. As a mapping on $\mathbb{R}^2$, $\boldsymbol{f}(x,y)=\begin{pmatrix} u(x,y)\\ v(x,y)\end{pmatrix}$. Thus show $d(\boldsymbol{f},B_r,\mathbf{0})=n$. **Hint:** You might consider $g(z)\equiv\prod_{j=1}^n(z-a_j)$ where the a_j are small real distinct numbers and argue that both this function and f are analytic but that $\mathbf{0}$ is a regular value for $\boldsymbol{g}$ although it is not so for $\boldsymbol{f}$. However, for each a_j small but distinct $d(\boldsymbol{f},B_r,\mathbf{0})=d(\boldsymbol{g},B_r,\mathbf{0})$.

17. Using Problem 16, prove the fundamental theorem of algebra as follows. Let $p(z)$ be a nonconstant polynomial of degree n, $p(z)=a_nz^n+a_{n-1}z^{n-1}+\cdots$ Show that for large enough r, $|p(z)|>|p(z)-a_nz^n|$ for all $z\in\partial B(0,r)$. Now from Problem 15 you can conclude $d(p,B_r,0)=d(f,B_r,0)=n$ where $f(z)=a_nz^n$.

18. Suppose $\boldsymbol{f}:\mathbb{R}^p\to\mathbb{R}^p$ satisfies $|\boldsymbol{f}(\boldsymbol{x})-\boldsymbol{f}(\boldsymbol{y})|\geq\alpha|\boldsymbol{x}-\boldsymbol{y}|$, $\alpha>0$. Show that $\boldsymbol{f}$ must map $\mathbb{R}^p$ onto $\mathbb{R}^p$. **Hint:** First show $\boldsymbol{f}$ is one to one. Then use invariance of domain. Next show, using the inequality, that the points not in $\boldsymbol{f}(\mathbb{R}^p)$ must form an open set because if $\boldsymbol{y}$ is such a point, then there can be no sequence $\{\boldsymbol{f}(\boldsymbol{x}_n)\}$ converging to it. Finally recall that $\mathbb{R}^p$ is connected.

19. Suppose D is a nonempty bounded open set in $\mathbb{R}^p$ and suppose $\boldsymbol{f} : D \to \partial D$ is continuous with $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{x}$ for $\boldsymbol{x} \in \partial D$. Show this cannot happen. **Hint:** Let $\boldsymbol{y} \in D$ and note that id and $\boldsymbol{f}$ agree on ∂D. Therefore, from properties of the degree, $d(\boldsymbol{f}, D, \boldsymbol{y}) = d(\mathrm{id}, D, \boldsymbol{y})$. Explain why this cannot occur.

20. Assume D is a closed ball in $\mathbb{R}^p$ and suppose $\boldsymbol{f} : D \to D$ is continuous. Use the above problem to conclude $\boldsymbol{f}$ has a fixed point. **Hint:** If no fixed point, let $\boldsymbol{g}(\boldsymbol{x})$ be the point on ∂D which results from extending the ray starting at $\boldsymbol{f}(\boldsymbol{x})$ to $\boldsymbol{x}$. This would be a continuous map from D to ∂D which does not move any point on ∂D. Draw a picture. This may be the easiest proof of the Brouwer fixed point theorem but note how dependent it is on the properties of the degree.

21. Use Corollary 15.6.9 to prove the invariance of domain theorem that if U is open and $\boldsymbol{f} : U \subseteq \mathbb{R}^p \to \mathbb{R}^p$ is continuous and one to one, then $\boldsymbol{f}(U)$ is open. This was discussed in the chapter but go through the details.

Chapter 16

Hausdorff Measure

16.1 Lipschitz Functions

Definition 16.1.1 *A function $\boldsymbol{f}: U \subseteq \mathbb{R}^n \to \mathbb{R}^m$ is Lipschitz if there is a constant K such that for all $x, y \in U$,*

$$|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{y})| \leq K |\boldsymbol{x} - \boldsymbol{y}|.$$

We assume $U \neq \emptyset$.

In what follows, dt will be used instead of dm in order to make the notation more familiar.

Lemma 16.1.2 *Suppose $f: [a,b] \to \mathbb{R}$ is Lipschitz continuous. Then f' exists a.e., is in $L^1([a,b])$, and*

$$f(x) = f(a) + \int_a^x f'(t)\, dt.$$

In fact, the almost everywhere existence of the derivative holds with only the assumption that f is increasing or of bounded variation on finite intervals. If $f: \mathbb{R} \to \mathbb{R}$ is Lipschitz, then f' is in $L^1_{loc}(\mathbb{R})$ and the above formula holds.

Proof: Since f is Lipschitz, it follows that it is absolutely continuous and so from Theorem 11.8.3, it has a derivative a.e. and one can recover the function from integrating the derivative as claimed. Also, from the formula for finding the derivative, it follows that if K is the Lipschitz constant, $|f'(t)| \leq K$. ■

Recall that it was shown earlier that the derivative of an increasing function exists a.e.(Theorem 9.13.4.)

16.2 Lipschitz Functions and Gateaux Derivatives

One can show that a Lipschitz continuous function defined on $\mathbb{R}^p$ has all Gateaux derivatives off some set of measure zero. Also, one can express $f(\boldsymbol{x}+\boldsymbol{v}) - f(\boldsymbol{x})$ as an appropriate one dimensional integral involving the Gateaux derivative in the direction $\boldsymbol{v}$.

Recall the Gateaux derivative is $D_{\boldsymbol{v}} f(\boldsymbol{x}) \equiv \lim_{h \to 0} \frac{f(\boldsymbol{x}+h\boldsymbol{v}) - f(\boldsymbol{x})}{h}$.

Corollary 16.2.1 *Suppose $f: \mathbb{R}^p \to \mathbb{R}$ is Lipschitz continuous,*

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq K |\boldsymbol{x} - \boldsymbol{y}|.$$

Then

$$f(\boldsymbol{x}+\boldsymbol{v}) - f(\boldsymbol{x}) = \int_0^1 D_{\boldsymbol{v}} f(\boldsymbol{x} + t\boldsymbol{v})\, dt$$

where the integrand is the Gateaux derivative and also $|D_{\boldsymbol{v}} f(\boldsymbol{x}+t\boldsymbol{v})| \leq K|\boldsymbol{v}|$ a.e.

Proof: $t \to f(\boldsymbol{x}+t\boldsymbol{v}) - f(\boldsymbol{x})$ is absolutely continuous because it is Lipschitz and so by Theorem 11.8.3 it is the integral of its derivative denoted by $g(t)$. But

$$g(t) = \lim_{h \to 0} \frac{1}{h} \int_t^{t+h} g(s)\, ds$$

$$= \lim_{h\to 0} \frac{f(\boldsymbol{x}+(t+h)\boldsymbol{v}) - f(\boldsymbol{x}) - (f(\boldsymbol{x}+t\boldsymbol{v}) - f(\boldsymbol{x}))}{h} \equiv D_{\boldsymbol{v}} f(\boldsymbol{x}+t\boldsymbol{v})$$

Also,

$$|D_{\boldsymbol{v}} f(\boldsymbol{x}+t\boldsymbol{v})| = \lim_{h\to 0} \left| \frac{f(\boldsymbol{x}+(t+h)\boldsymbol{v}) - f(\boldsymbol{x}+t\boldsymbol{v})}{h} \right| \leq \frac{K|h||\boldsymbol{v}|}{|h|} = K|\boldsymbol{v}| \ \blacksquare$$

Next I want to verify that off a set of measure zero, the Gateaux derivative exists. The above essentially does it but maybe is not completely clear because it only works along one dimensional lines.

Proposition 16.2.2 *Let $f : \mathbb{R}^p \to \mathbb{R}$ be Lipschitz continuous with constant K. Then there is a set of measure zero N such that if $\boldsymbol{x} \notin N$, then $D_{\boldsymbol{v}} f(\boldsymbol{x})$ exists. Also, $D_{\boldsymbol{v}} f(\boldsymbol{x}) = D_{\boldsymbol{v}/|\boldsymbol{v}|} f(\boldsymbol{x})|\boldsymbol{v}|$ so $|D_{\boldsymbol{v}} f(\boldsymbol{x})| \leq K|\boldsymbol{v}|$ a.e. $\boldsymbol{x}$.*

Proof: Since $x_k \to f(x_1, \cdots, x_k, \cdots, x_p)$ is Lipschitz continuous, it follows from the one variable considerations above that

$$D_{\boldsymbol{e}_k} f(x_1, \cdots, x_k, \cdots, x_p) \equiv f_{x_k}(x_1, \cdots, x_k, \cdots, x_p)$$

exists for a.e. x_k. If N is the set where this partial derivative does not exist, then it is a Borel set because f is continuous. Hence

$$m_p(N) = \int_{\mathbb{R}^{p-1}} \int_{\mathbb{R}} \mathscr{X}_N(x_1, \cdots, x_k, \cdots, x_p)\, dx_k d\hat{\boldsymbol{x}}_k$$

where $\hat{\boldsymbol{x}}_k$ refers to $(x_1, \cdots, x_{k-1}, x_{k+1} \cdots, x_p) \in \mathbb{R}^{p-1}$. Thus $m_p(N) = 0$. It follows that off a set of measure zero, all partial derivatives exist. The next task is to show that $D_{\boldsymbol{v}} f(\boldsymbol{x})$ exists off some other larger set of measure zero.

$t \to f(\boldsymbol{x}+t\boldsymbol{v})$ is Lipschitz for each $\boldsymbol{v}$ with $|\boldsymbol{v}| = 1$. Let N be the set where the directional derivative does not exist. Let U be an orthogonal matrix with $\boldsymbol{v} = U\boldsymbol{e}_1$. Let $g(\boldsymbol{x}) \equiv f(U\boldsymbol{x})$. Then $g(\boldsymbol{x}+t\boldsymbol{e}_1) = f(U\boldsymbol{x}+tU\boldsymbol{e}_1) = f(U\boldsymbol{x}+t\boldsymbol{v})$. Thus $\boldsymbol{x} \in N$ if and only if $U\boldsymbol{x}$ is in the set where g_{x_1} fails to exist. By what was just shown, this set where g_{x_1} does not exist has measure zero and since $\det(U) = 1$, it follows from Theorem 11.13.2 that the set where the directional derivative for f in the direction $\boldsymbol{v}$ fails to exist also has measure zero. Thus if D is a dense countable subset of S^{p-1}, there exists a set of measure zero N, the union of those just described, such that for all $\boldsymbol{x} \notin N$ and any $\boldsymbol{v} \in D$,

$$\lim_{h\to 0} \frac{f(\boldsymbol{x}+h\boldsymbol{v}) - f(\boldsymbol{x})}{h} \equiv D_{\boldsymbol{v}} f(\boldsymbol{x})$$

This will imply that the directional derivative will exist for all $\boldsymbol{v} \in S^{p-1}$ provided $\boldsymbol{x} \notin N$. So pick $\boldsymbol{v} \in S^{p-1}$ the unit sphere. By continuity of f,

$$\boldsymbol{x} \to \liminf_{h\to 0} \frac{f(\boldsymbol{x}+h\boldsymbol{v}) - f(\boldsymbol{x})}{h} \text{ and } \boldsymbol{x} \to \limsup_{h\to 0} \frac{f(\boldsymbol{x}+h\boldsymbol{v}) - f(\boldsymbol{x})}{h}$$

are both measurable.

$$\left| \frac{f(\boldsymbol{x}+h\boldsymbol{v}) - f(\boldsymbol{x})}{h} - \frac{f(\boldsymbol{x}+h\hat{\boldsymbol{v}}) - f(\boldsymbol{x})}{h} \right| \leq K|\boldsymbol{v}-\hat{\boldsymbol{v}}| < \varepsilon$$

if $\hat{\boldsymbol{v}} \in D$ is chosen close enough to $\boldsymbol{v}$. So pick such a $\hat{\boldsymbol{v}}$. Letting $\Delta(h) \equiv \frac{f(\boldsymbol{x}+h\boldsymbol{v})-f(\boldsymbol{x})}{h}$ and $\hat{\Delta}(h) \equiv \frac{f(\boldsymbol{x}+h\hat{\boldsymbol{v}})-f(\boldsymbol{x})}{h}$,

$$\limsup_{h\to 0} \frac{f(\boldsymbol{x}+h\boldsymbol{v})-f(\boldsymbol{x})}{h} - \liminf_{h\to 0} \frac{f(\boldsymbol{x}+h\boldsymbol{v})-f(\boldsymbol{x})}{h} =$$

$$\begin{aligned} & \limsup_{h\to 0} \left(\Delta(h)-\hat{\Delta}(h)+\hat{\Delta}(h)\right) - \liminf_{h\to 0} \left(\Delta(h)-\hat{\Delta}(h)+\hat{\Delta}(h)\right) \\ \leq\ & 2\varepsilon + \limsup_{h\to 0} \hat{\Delta}(h) - \liminf_{h\to 0} \hat{\Delta}(h) = 2\varepsilon \end{aligned}$$

Since ε is arbitrary, $\limsup_{h\to 0} \frac{f(\boldsymbol{x}+h\boldsymbol{v})-f(\boldsymbol{x})}{h} - \liminf_{h\to 0} \frac{f(\boldsymbol{x}+h\boldsymbol{v})-f(\boldsymbol{x})}{h} = 0$. Now if $\boldsymbol{x} \notin N$ and $\boldsymbol{v}$ is totally arbitrary and without loss of generality not zero,

$$\frac{f(\boldsymbol{x}+h\boldsymbol{v})-f(\boldsymbol{x})}{h} = \frac{f\left(\boldsymbol{x}+h|\boldsymbol{v}|\frac{\boldsymbol{v}}{|\boldsymbol{v}|}\right)-f(\boldsymbol{x})}{|\boldsymbol{v}|h}|\boldsymbol{v}| \to D_{\boldsymbol{v}/|\boldsymbol{v}|}f(\boldsymbol{x})|\boldsymbol{v}|$$

Since the limit of the difference quotient exists, this shows that if $\boldsymbol{x} \notin N$,

$$D_{\boldsymbol{v}}f(\boldsymbol{x}) \equiv \lim_{h\to 0} \frac{f(\boldsymbol{x}+h\boldsymbol{v})-f(\boldsymbol{x})}{h} = D_{\boldsymbol{v}/|\boldsymbol{v}|}f(\boldsymbol{x})|\boldsymbol{v}|$$

The claim about the size of the directional derivative follows right away because

$$\left|\frac{f(\boldsymbol{x}+h\boldsymbol{v})-f(\boldsymbol{x})}{h}\right| \leq \frac{K|h||\boldsymbol{v}|}{|h|} = K|\boldsymbol{v}|. \blacksquare$$

Of course this has not proved Rademacher's theorem which says that $Df(\boldsymbol{x})$ exists a.e. $\boldsymbol{x}$ because it only gives existence of directional derivatives in every direction. This existence of directional derivatives is a weaker condition as shown in Problem 12 on Page 207 or some other problems near this one.

16.3 Rademacher's Theorem

It turns out that Lipschitz functions on $\mathbb{R}^p$ can be differentiated a.e. This is called Rademacher's theorem. The following lemma comes from Corollary 16.2.1 and Proposition 16.2.2

Lemma 16.3.1 *Let $f : \mathbb{R}^p \to \mathbb{R}$ be Lipschitz with constant K. Then*

$$f(x_1,\cdots,x_k+h,\cdots,x_p) - f(x_1,\cdots,x_k,\cdots,x_p) = \int_0^1 f_{x_k}(x_1,\cdots,x_k+th,\cdots,x_p)\,h dt$$

Then we obtain the following corollary.

Corollary 16.3.2 *Let $f : \mathbb{R}^p \to \mathbb{R}$ be Lipschitz with constant K. Then for*

$$\boldsymbol{x} \equiv (x_1,\cdots,x_p), \boldsymbol{v} \equiv (v_1,\cdots,v_p)$$

$$f(\boldsymbol{x}+\boldsymbol{v}) - f(\boldsymbol{x}) = \sum_{k=1}^{p} \int_0^1 f_{x_k}(x_1+v_1,\cdots,x_{k-1}+v_{k-1},x_k+tv_k,\cdots,x_p)\,v_k dt \tag{16.1}$$

With this corollary, here is Rademacher's theorem.

Theorem 16.3.3 *Let f be Lipschitz on $\mathbb{R}^p$. Then $Df(\boldsymbol{x})$ exists for a.e. $\boldsymbol{x}$.*

Proof: f is Lipschitz, so f is bounded on every ball. From 16.1, and $\boldsymbol{x} \equiv (x_1, \cdots, x_p)$, similar for $\boldsymbol{v}$,

$$f(\boldsymbol{x}+\boldsymbol{v}) - (f(\boldsymbol{x}) + \nabla f(\boldsymbol{x}) \cdot \boldsymbol{v}) =$$

$$\sum_{k=1}^{p} \int_0^1 f_{x_k}(x_1+v_1, \cdots, x_{k-1}+v_{k-1}, x_k+tv_k, \cdots, x_p) v_k - f_{x_k}(\boldsymbol{x}) v_k dt \tag{16.2}$$

Now every function in sight is Borel measurable and in L^1_{loc} so m_p a.e. point is a Lebesgue point for every function mentioned above thanks to Corollary 11.3.4. Letting $\boldsymbol{x}$ be such a Lebesgue point,

$$\frac{1}{m_p(\boldsymbol{x}, r)} \int_{B(\boldsymbol{x},r)} f(\boldsymbol{y}+\boldsymbol{v}) - (f(\boldsymbol{y}) + \nabla f(\boldsymbol{y}) \cdot \boldsymbol{v}) \, dy =$$

$$\sum_{k=1}^{p} \int_0^1 \frac{1}{m_p(\boldsymbol{x}, r)} \int_{B(\boldsymbol{x},r)} f_{x_k}(y_1+v_1, \cdots, y_{k-1}+v_{k-1}, y_k+tv_k, \cdots, y_p) v_k - f_{x_k}(\boldsymbol{y}) v_k dy dt$$

Let the last term be e_r. Since $\boldsymbol{x}$ is a Lebesgue point, for given $\varepsilon > 0$, $|e_r| < (\varepsilon/2)|\boldsymbol{v}|$ whenever r is small enough and $|\boldsymbol{v}|$ is small enough. Picking such an r, and $\boldsymbol{v}$ sufficiently small,

$$\frac{1}{m_p(\boldsymbol{x}, r)} \int_{B(\boldsymbol{x},r)} f(\boldsymbol{y}+\boldsymbol{v}) - (f(\boldsymbol{y}) + \nabla f(\boldsymbol{y}) \cdot \boldsymbol{v}) \, dy < \frac{\varepsilon}{2} |\boldsymbol{v}|$$

Now let $r \to 0$ to obtain that whenever $|\boldsymbol{v}|$ is small enough,

$$f(\boldsymbol{x}+\boldsymbol{v}) - (f(\boldsymbol{x}) + \nabla f(\boldsymbol{x}) \cdot \boldsymbol{v}) \leq \frac{\varepsilon}{2} |\boldsymbol{v}| < \varepsilon |\boldsymbol{v}|$$

showing that at every Lebesgue point of f and its partial derivatives, the right side of 16.2 is $o(|\boldsymbol{v}|)$. ■

For $|\boldsymbol{w}| = 1$, denote the measure of Section 11.17 defined on the unit sphere S^{p-1} as σ. Let $N_{\boldsymbol{w}}$ be defined as those $t \in [0, \infty)$ for which $D_{\boldsymbol{w}} u(\boldsymbol{x}+t\boldsymbol{w}) \neq \nabla u(\boldsymbol{x}+t\boldsymbol{w}) \cdot \boldsymbol{w}$.

$$B \equiv \left\{ \boldsymbol{w} \in S^{p-1} : N_{\boldsymbol{w}} \text{ has positive measure} \right\}$$

This is contained in the set of points of $\mathbb{R}^p$ where $v(\cdot) \equiv u(\boldsymbol{x}+\cdot)$ is not differentiable. Thus from Section 11.17 the measure of this set is $\int_B \int_{N_{\boldsymbol{w}}} \rho^{n-1} d\rho d\sigma(w)$. This must equal zero from what was just shown about the derivative of the Lipschitz function v existing a.e. and so $\sigma(B) = 0$. The claimed formula follows from this. Thus we obtain the following corollary.

Corollary 16.3.4 *Let u be Lipschitz. Then for any $\boldsymbol{x}$ and $\boldsymbol{v} \in S^{p-1} \setminus B_{\boldsymbol{x}}$ where $\sigma(B_{\boldsymbol{x}}) = 0$, it follows that for all t,*

$$u(\boldsymbol{x}+t\boldsymbol{v}) - u(\boldsymbol{x}) = \int_0^t D_{\boldsymbol{v}} u(\boldsymbol{x}+s\boldsymbol{v}) \, ds = \int_0^t \nabla u(\boldsymbol{x}+s\boldsymbol{v}) \cdot \boldsymbol{v} ds$$

In all of the above, the function u is defined on all of $\mathbb{R}^p$. However, it is always the case that Lipschitz functions can be extended off a given set. Thus if a Lipschitz function is defined on some set Ω, then it can always be considered the restriction to Ω of a Lipschitz map defined on all of $\mathbb{R}^p$.

Theorem 16.3.5 *If $\boldsymbol{h}:\Omega\to\mathbb{R}^m$ is Lipschitz, then there exists $\overline{\boldsymbol{h}}:\mathbb{R}^p\to\mathbb{R}^m$ which extends $\boldsymbol{h}$ and is also Lipschitz.*

Proof: It suffices to assume $m=1$ because if this is shown, it may be applied to the components of $\boldsymbol{h}$ to get the desired result. Suppose

$$|h(\boldsymbol{x})-h(\boldsymbol{y})|\leq K|\boldsymbol{x}-\boldsymbol{y}|. \tag{16.3}$$

Define

$$\overline{h}(\boldsymbol{x})\equiv\inf\{h(\boldsymbol{w})+K|\boldsymbol{x}-\boldsymbol{w}|:\boldsymbol{w}\in\Omega\}. \tag{16.4}$$

If $\boldsymbol{x}\in\Omega$, then for all $\boldsymbol{w}\in\Omega$,

$$h(\boldsymbol{w})+K|\boldsymbol{x}-\boldsymbol{w}|\geq h(\boldsymbol{x})$$

by 16.3. This shows $h(\boldsymbol{x})\leq\overline{h}(\boldsymbol{x})$. But also you could take $\boldsymbol{w}=\boldsymbol{x}$ in 16.4 which yields $\overline{h}(\boldsymbol{x})\leq h(\boldsymbol{x})$. Therefore $\overline{h}(\boldsymbol{x})=h(\boldsymbol{x})$ if $\boldsymbol{x}\in\Omega$.

Now suppose $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^p$ and consider $\left|\overline{h}(\boldsymbol{x})-\overline{h}(\boldsymbol{y})\right|$. Without loss of generality assume $\overline{h}(\boldsymbol{x})\geq\overline{h}(\boldsymbol{y})$. (If not, repeat the following argument with $\boldsymbol{x}$ and $\boldsymbol{y}$ interchanged.) Pick $\boldsymbol{w}\in\Omega$ such that

$$h(\boldsymbol{w})+K|\boldsymbol{y}-\boldsymbol{w}|-\varepsilon<\overline{h}(\boldsymbol{y}).$$

Then

$$\left|\overline{h}(\boldsymbol{x})-\overline{h}(\boldsymbol{y})\right|=\overline{h}(\boldsymbol{x})-\overline{h}(\boldsymbol{y})\leq h(\boldsymbol{w})+K|\boldsymbol{x}-\boldsymbol{w}|-$$
$$[h(\boldsymbol{w})+K|\boldsymbol{y}-\boldsymbol{w}|-\varepsilon]\leq K|\boldsymbol{x}-\boldsymbol{y}|+\varepsilon.$$

Since ε is arbitrary, $\left|\overline{h}(\boldsymbol{x})-\overline{h}(\boldsymbol{y})\right|\leq K|\boldsymbol{x}-\boldsymbol{y}|$. ■

16.4 Approximating Lipschitz Maps

A very useful idea is about approximating Lipschitz functions with smooth ones.

Lemma 16.4.1 *Let $u:\mathbb{R}^p\to\mathbb{R}$ be Lipschitz with Lipschitz constant K. Let*

$$u_n\equiv u*\phi_n\equiv\int u(\boldsymbol{x}-\boldsymbol{y})\,dm_p(\boldsymbol{y})$$

where $\{\phi_n\}$ is a mollifier,

$$\phi_n(\boldsymbol{y})\equiv n^p\phi(n\boldsymbol{y}),\ \int\phi(\boldsymbol{y})\,dm_p(\boldsymbol{y})=1,\ \phi(\boldsymbol{y})\geq 0,\ \phi\in C_c^\infty(B(\boldsymbol{0},1))$$

Then

$$\nabla u_n(\boldsymbol{x})=\nabla u*\phi_n(\boldsymbol{x}) \tag{16.5}$$

where ∇u is defined almost everywhere according to Proposition 16.2.2. In fact,

$$\int_a^b\frac{\partial u}{\partial x_i}(\boldsymbol{x}+t\boldsymbol{e}_i)\,dt=u(\boldsymbol{x}+b\boldsymbol{e}_i)-u(\boldsymbol{x}+a\boldsymbol{e}_i) \tag{16.6}$$

and $\left|\frac{\partial u}{\partial x_i}\right|\leq K$ so $|\nabla u(\boldsymbol{x})|\leq\sqrt{p}K$ for a.e. $\boldsymbol{x}$. Also, $u_n(\boldsymbol{x})\to u(\boldsymbol{x})$ uniformly on $\mathbb{R}^p$ and for a suitable subsequence, still denoted with n, $\nabla u_n(\boldsymbol{x})\to\nabla u(\boldsymbol{x})$ for a.e. $\boldsymbol{x}$. Also if U is a bounded set, then the components of ∇u_n converge to the components of u in $L^1(U)$. If, in addition to this, u is in $C^1(\bar{U})$ for U a bounded open set, then $\nabla u_n(\boldsymbol{x})\to\nabla u(\boldsymbol{x})$ uniformly for all $\boldsymbol{x}$.

Proof: To get the existence of the gradient satisfying the condition given in 16.6, apply Proposition 16.2.2 to each variable. Now

$$\begin{aligned}\frac{u_n(\boldsymbol{x}+h\boldsymbol{e}_i)-u_n(\boldsymbol{x})}{h} &= \int_{\mathbb{R}^p}\left(\frac{u(\boldsymbol{x}+h\boldsymbol{e}_i-\boldsymbol{y})-u(\boldsymbol{x}-\boldsymbol{y})}{h}\right)\phi_n(\boldsymbol{y})\,dm_p(\boldsymbol{y})\\ &= \int_{B(\mathbf{0},\frac{1}{n})}\left(\frac{u(\boldsymbol{x}+h\boldsymbol{e}_i-\boldsymbol{y})-u(\boldsymbol{x}-\boldsymbol{y})}{h}\right)\phi_n(\boldsymbol{y})\,dm_p(\boldsymbol{y})\\ &= \int_{B(\mathbf{0},1)}\left(\frac{u(\boldsymbol{x}+h\boldsymbol{e}_i-\boldsymbol{y})-u(\boldsymbol{x}-\boldsymbol{y})}{h}\right)\phi_n(\boldsymbol{y})\,dm_p(\boldsymbol{y})\end{aligned}$$

Now if $\boldsymbol{x}-\boldsymbol{y}$ is off a set of measure zero, the above difference quotient converges to $\frac{\partial u}{\partial x_i}(\boldsymbol{x}-\boldsymbol{y})$. You just use Proposition 16.2.2 on the i^{th} variable. If h_k is any sequence converging to 0, you can apply the dominated conergence theorem in the above and obtain

$$\frac{\partial u_n(\boldsymbol{x})}{\partial x_i}=\int_{B(\mathbf{0},1)}\frac{\partial u(\boldsymbol{x}-\boldsymbol{y})}{\partial x_i}\phi_n(\boldsymbol{y})\,dm_p(\boldsymbol{y})=\frac{\partial u}{\partial x_i}*\phi_n(\boldsymbol{x})$$

This proves 16.5.

$$\begin{aligned}|u_n(\boldsymbol{x})-u(\boldsymbol{x})| &\le \int_{\mathbb{R}^p}|u(\boldsymbol{x}-\boldsymbol{y})-u(\boldsymbol{x})|\,\phi_n(\boldsymbol{y})\,dm_p(\boldsymbol{y})\\ &= \int_{B(\mathbf{0},\frac{1}{n})}|u(\boldsymbol{x}-\boldsymbol{y})-u(\boldsymbol{x})|\,\phi_n(\boldsymbol{y})\,dm_p(\boldsymbol{y})\end{aligned}$$

by uniform continuity of u coming from the Lipschitz condition, when n is large enough, this is no larger than $\int_{\mathbb{R}^p}\varepsilon\phi_n(\boldsymbol{y})\,dm_p(\boldsymbol{y})=\varepsilon$ and so uniform convergence holds.

Now consider the last claim. From the first part,

$$\begin{aligned}|u_{nx_i}(\boldsymbol{x})-u_{x_i}(\boldsymbol{x})| &= \left|\int_{B(\mathbf{0},\frac{1}{n})}u_{x_i}(\boldsymbol{x}-\boldsymbol{y})\,\phi_n(\boldsymbol{y})\,dm_p(\boldsymbol{y})-u_{x_i}(\boldsymbol{x})\right|\\ &= \left|\int_{B(\boldsymbol{x},\frac{1}{n})}u_{x_i}(\boldsymbol{z})\,\phi_n(\boldsymbol{x}-\boldsymbol{z})\,dm_p(\boldsymbol{z})-u_{x_i}(\boldsymbol{x})\right|\end{aligned}$$

$$|u_{nx_i}(\boldsymbol{x})-u_{x_i}(\boldsymbol{x})|\le\int_{\mathbb{R}^p}|u_{x_i}(\boldsymbol{x}-\boldsymbol{y})-u_{x_i}(\boldsymbol{x})|\,\phi_n(\boldsymbol{y})\,dm_p(\boldsymbol{y})$$

$$\begin{aligned}&= \int_{B(\mathbf{0},\frac{1}{n})}|u_{x_i}(\boldsymbol{x}-\boldsymbol{y})-u_{x_i}(\boldsymbol{x})|\,\phi_n(\boldsymbol{y})\,dm_p(\boldsymbol{y})\\ &= \int_{B(\boldsymbol{x},\frac{1}{n})}|u_{x_i}(\boldsymbol{z})-u_{x_i}(\boldsymbol{x})|\,\phi_n(\boldsymbol{x}-\boldsymbol{z})\,dm_p(\boldsymbol{z})\end{aligned}$$

$$\begin{aligned}&\le n^p\int_{B(\boldsymbol{x},\frac{1}{n})}|u_{x_i}(\boldsymbol{z})-u_{x_i}(\boldsymbol{x})|\,\phi(n(\boldsymbol{x}-\boldsymbol{z}))\,dm_p(\boldsymbol{z})\\ &\le \frac{C}{m_p\left(B\left(\mathbf{0},\frac{1}{n}\right)\right)}\int_{B(\boldsymbol{x},\frac{1}{n})}|u_{x_i}(\boldsymbol{z})-u_{x_i}(\boldsymbol{x})|\,dm_p(\boldsymbol{z})\end{aligned}$$

which converges to 0 for a.e. x, in fact at any Lebesgue point. This is because u_{x_i} is bounded by K and so is in L^1_{loc}. Convergence in $L^1(U)$ follows from this and the dominated convergence theorem. Indeed,

$$\left|\frac{u_n(x+he_i)-u_n(x)}{h}\right| = \int_{\mathbb{R}^p}\left|\frac{u(x+he_i-y)-u(x-y)}{h}\right|\phi_n(y)\,dm_p(y) \le K$$

so the partial derivatives of u_n are bounded by K.

It only remains to verify the last claim. However, ∇u is uniformly continuous and ∇u_n equals $\nabla u * \phi_n$ so it is clearly true that convergence takes place uniformly. ■

Note that the above holds just as well if u has values in some $\mathbb{R}^m$ and the same proof would work, replacing $|\cdot|$ with $\|\cdot\|$ or the Euclidean norm $|\cdot|$.

16.5 Weak Derivatives

A related concept is that of weak derivatives. Let $\Omega \subseteq \mathbb{R}^p$ be an open set. A distribution on Ω is defined to be a linear functional on $C_c^\infty(\Omega)$, called the space of test functions. The space of all such linear functionals will be denoted by $\mathscr{D}^*(\Omega)$. Actually, more is sometimes done here. One imposes a topology on $C_c^\infty(\Omega)$ making it into a topological vector space, and when this has been done, $\mathscr{D}'(\Omega)$ is defined as the continuous linear maps. To see this, consult the book by Yosida [59] or the book by Rudin [50]. I am ignoring this topology because in practice, one is usually more interested in some other topology which is much less exotic. Thus $\mathscr{D}^*(\Omega)$ is an algebraic dual which has nothing to do with topology.

The following is a basic lemma which will be used in what follows. First recall the following definition.

Definition 16.5.1 *For Ω an open set in $\mathbb{R}^n$, $C_c^\infty(\Omega)$ denotes those functions ϕ which are infinitely differentiable and have compact support in Ω. This is a nonempty set of functions by Lemma 12.5.3.*

With this definition, the fundamental lemma is as follows.

Lemma 16.5.2 *Suppose $f \in L^1_{loc}(\mathbb{R}^n)$ and suppose $\int f\phi dx = 0$ for all $\phi \in C_c^\infty(\mathbb{R}^n)$. Then $f(x) = 0$ a.e. x.*

Proof: Without loss of generality f is real-valued. Let $E \equiv \{x : f(x) > \varepsilon\}$ and let $E_m \equiv E \cap B(0,m)$. We show that $m(E_m) = 0$. If not, there exists an open set V, and a compact set K satisfying

$$K \subseteq E_m \subseteq V \subseteq B(0,m),\ m_p(V\setminus K) < 4^{-1}m(E_m),$$

$$\int_{V\setminus K}|f|\,dx < \varepsilon 4^{-1}m_p(E_m).$$

Let H and W be open sets satisfying $K \subseteq H \subseteq \overline{H} \subseteq W \subseteq \overline{W} \subseteq V$ and let $\overline{H} \prec g \prec W$ where the symbol, $\prec$, has the same meaning as it does in Definition 3.12.3. That is, g equals 1 on $\overline{H}$ and has compact support contained in W. Then let ϕ_δ be a mollifier and let $h \equiv g * \phi_\delta$ for δ small enough that $K \prec h \prec V$. Thus

$$\begin{aligned} 0 &= \int fh\,dx = \int_K f\,dx + \int_{V\setminus K} fh\,dx \\ &\ge \varepsilon m_p(K) - \varepsilon 4^{-1}m_p(E_m) \\ &\ge \varepsilon\left(m_p(E_m) - 4^{-1}m_p(E_m)\right) - \varepsilon 4^{-1}m_p(E_m) \ge 2^{-1}\varepsilon m_p(E_m). \end{aligned}$$

Therefore, $m_p(E_m) = 0$, a contradiction. Thus $m_p(E) \leq \sum_{m=1}^{\infty} m_p(E_m) = 0$ and so, since $\varepsilon > 0$ is arbitrary, $m_p(\{x : f(x) > 0\}) = 0$. Similarly $m(\{x : f(x) < 0\}) = 0$. If f is complex valued, the above applies to the real and imaginary parts. ■

Example: The space $L^1_{loc}(\Omega)$ may be considered as a subset of $\mathscr{D}^*(\Omega)$ as follows. $f(\phi) \equiv \int_\Omega f(x)\phi(x)dx$ for all $\phi \in C_c^\infty(\Omega)$. Recall that $f \in L^1_{loc}(\Omega)$ if $f\mathscr{X}_K \in L^1(\Omega)$ whenever K is compact.

This is well defined thanks to Lemma 16.5.2.

Example: $\delta_x \in \mathscr{D}^*(\Omega)$ where $\delta_x(\phi) \equiv \phi(x)$.

It will be observed from the above two examples and a little thought that $\mathscr{D}^*(\Omega)$ is truly enormous. We shall define the derivative of a distribution in such a way that it agrees with the usual notion of a derivative on those distributions which are also continuously differentiable functions. With this in mind, let f be the restriction to the open set Ω of a smooth function defined on $\mathbb{R}^p$. Then $D_{x_i} f$ makes sense and for $\phi \in C_c^\infty(\Omega)$

$$D_{x_i} f(\phi) \equiv \int_\Omega D_{x_i} f(x)\phi(x)dx = -\int_\Omega f D_{x_i}\phi dx = -f(D_{x_i}\phi).$$

Motivated by this, here is the definition of a weak derivative.

Definition 16.5.3

For $T \in \mathscr{D}^*(\Omega)$, $D_{x_i}T(\phi) \equiv -T(D_{x_i}\phi)$.

One can continue taking derivatives indefinitely. Thus, $D_{x_i x_j}T \equiv D_{x_i}\left(D_{x_j}T\right)$ and it is clear that all mixed partial derivatives are equal because this holds for the functions in $C_c^\infty(\Omega)$. Thus one can differentiate virtually anything, even functions that may be discontinuous everywhere. However the notion of "derivative" is very weak, hence the name, "weak derivatives".

Example: Let $\Omega = \mathbb{R}$ and let

$$H(x) \equiv \begin{cases} 1 \text{ if } x \geq 0, \\ 0 \text{ if } x < 0. \end{cases}$$

Then $DH(\phi) = -\int H(x)\phi'(x)dx = \phi(0) = \delta_0(\phi)$. Note that in this example, DH is not a function.

What happens when Df is a function?

Theorem 16.5.4

Let $\Omega = (a,b)$ and suppose that f and Df are both in $L^1(a,b)$. Then f is equal to a continuous function a.e., still denoted by f and $f(x) = f(a) + \int_a^x Df(t)dt$.

Proof: Consider $f - \int_a^{(\cdot)} Df(t)dt \equiv T$. Is it equal to some constant a.e.? Let $\phi \in C_c^\infty(a,b)$. By Fubini's theorem,

$$\begin{aligned} DT(\phi) &\equiv \int_a^b \left(f(x) - \int_a^x Df(t)dt\right)\phi'(x)dx \\ &= \int_a^b f(x)\phi'(x)dx - \int_a^b \int_a^x Df(t)\phi'(x)dtdx \end{aligned}$$

$$\begin{aligned} &\equiv -Df(\phi) - \int_a^b Df(t)\int_t^b \phi'(x)dxdt = -Df(\phi) + \int_a^b Df(t)\phi(t)dt \\ &= -Df(\phi) + Df(\phi) = 0 \end{aligned}$$

Thus the theorem is proved if it is shown that whenever $DT = 0$, it follows that T is a constant. This is the following lemma.

Lemma 16.5.5 *Let $T \in \mathscr{D}^*(a,b)$ and suppose $DT = 0$. Then there exists a constant C such that $T(\phi) = \int_a^b C\phi dx$.*

Proof: $T(D\phi) = 0$ for all $\phi \in C_c^\infty(a,b)$ from the definition of $DT = 0$. Let $\phi_0 \in C_c^\infty(a,b)$, $\int_a^b \phi_0(x)\,dx = 1$, and let

$$\psi_\phi(x) = \int_a^x [\phi(t) - \left(\int_a^b \phi(y)\,dy\right)\phi_0(t)]dt$$

for $\phi \in C_c^\infty(a,b)$. Thus $\psi_\phi \in C_c^\infty(a,b)$ and $D\psi_\phi = \phi - \left(\int_a^b \phi(y)\,dy\right)\phi_0$. Therefore, $\phi = D\psi_\phi + \left(\int_a^b \phi(y)\,dy\right)\phi_0$ and so

$$T(\phi) = T(D\psi_\phi) + \left(\int_a^b \phi(y)\,dy\right)T(\phi_0) = \int_a^b T(\phi_0)\,\phi(y)\,dy.$$

Let $C = T\phi_0$. ■

It follows from this lemma that $f(x) - \int_a^x Df(t)\,dt = C$ for some constant which we denote as $f(a)$ so that $f(x) = f(a) + \int_a^x Df(t)\,dt$. ■

Theorem 16.5.4 says that $f(x) = f(a) + \int_a^x Df(t)\,dt$ whenever it makes sense to write $\int_a^x Df(t)\,dt$, if Df is interpreted as a weak derivative. Somehow, this is the way it ought to be. It follows from the fundamental theorem of calculus that $f'(x)$ exists for a.e. x in the classical sense where the derivative is taken in the sense of a limit of difference quotients and $f'(x) = Df(x)$. This raises an interesting question. Suppose f is continuous on $[a,b]$ and $f'(x)$ exists in the classical sense for a.e. x. Does it follow that $f(x) = f(a) + \int_a^x f'(t)\,dt$? The answer is no. You can build such an example from the Cantor function which is increasing and has a derivative a.e. which equals 0 a.e. and yet climbs from 0 to 1, Problem 4 on Page 273. Thus, in a sense weak derivatives are more agreeable than the classical ones.

16.6 Definition of Hausdorff Measures

First I will discuss some outer measures. In all that is done here, $\alpha(p)$ will be the volume of the ball in $\mathbb{R}^p$ which has radius 1. Hausdorff measures are very geometrically motivated and so the norm in $\mathbb{R}^p$ will be the usual Euclidean norm unless indicated otherwise.

Definition 16.6.1 *For a set E, denote by $r(E)$ the number which is half the diameter of E. Thus*

$$r(E) \equiv \frac{1}{2}\sup\{|\boldsymbol{x}-\boldsymbol{y}| : \boldsymbol{x},\boldsymbol{y} \in E\} \equiv \frac{1}{2}\operatorname{diam}(E)$$

Let $E \subseteq \mathbb{R}^p$. $\mathscr{H}_\delta^s(E) \equiv \inf\{\sum_{j=1}^\infty \beta(s)(r(C_j))^s : E \subseteq \cup_{j=1}^\infty C_j, r(C_j) < \delta\}$. Define $\mathscr{H}^s(E) \equiv \lim_{\delta\to 0+}\mathscr{H}_\delta^s(E)$. In case $s = 0$ and E is an infinite set, then for small enough δ, $\mathscr{H}_\delta^0(E) > m$ for any positive m so $\mathscr{H}^0(E) = \infty$ and in case E is a finite set, then $\mathscr{H}_\delta^0(E)$ will clearly be the number of things in E for all δ small enough. Thus $\mathscr{H}^0$ is just counting measure and the $\mathscr{H}_\delta^0$ are outer measures converging to $\mathscr{H}^0$ provided we define $\mathscr{H}^0(\emptyset) \equiv 0$.

Note that $\mathcal{H}^s_\delta(E)$ if you make δ smaller, $\mathcal{H}^s_\delta(E)$ will become larger and so the limit clearly exists.

In the above definition, $\beta(s)$ is an appropriate **positive constant** depending on s. It will turn out that for p an integer, $\beta(p) = \alpha(p)$ where $\alpha(p)$ is the Lebesgue measure of the unit ball, $B(\mathbf{0},1)$ where the Euclidean norm is used to determine this ball.

Lemma 16.6.2 *$\mathcal{H}^s$ and $\mathcal{H}^s_\delta$ are outer measures for all $s \geq 0$.*

Proof: The case $s = 0$ comes directly from the definition so assume $s > 0$. If $A \subseteq B$, then $\mathcal{H}^s(A) \leq \mathcal{H}^s(B)$ with similar assertions valid for $\mathcal{H}^s_\delta$. To see that $\mathcal{H}^s_\delta(\emptyset) = 0$, let $C_j \equiv B\left(\mathbf{0}, \varepsilon^{(1/s)} 2^{-(j+1)/s}\right)$ where $\varepsilon^{(1/s)} < \delta$ so that $\mathcal{H}^s_\delta(\emptyset) \leq \sum_{j=1}^\infty \beta(s) \left(\varepsilon^{1/s} 2^{-(j+1)/s}\right)^s = \beta(s)\frac{\varepsilon}{2}$. Since ε is arbitrary, $\mathcal{H}^s(\emptyset) = 0$.

Suppose $E = \cup_{i=1}^\infty E_i$ and $\mathcal{H}^s_\delta(E_i) < \infty$ for each i. Let $\{C^i_j\}_{j=1}^\infty$ be a covering of E_i with $\sum_{j=1}^\infty \beta(s)(r(C^i_j))^s - \varepsilon/2^i < \mathcal{H}^s_\delta(E_i)$ and $\operatorname{diam}(C^i_j) \leq \delta$. Then

$$\mathcal{H}^s_\delta(E) \leq \sum_{i=1}^\infty \sum_{j=1}^\infty \beta(s)(r(C^i_j))^s \leq \sum_{i=1}^\infty \mathcal{H}^s_\delta(E_i) + \varepsilon/2^i \leq \varepsilon + \sum_{i=1}^\infty \mathcal{H}^s_\delta(E_i).$$

It follows that since $\varepsilon > 0$ is arbitrary, $\mathcal{H}^s_\delta(E) \leq \sum_{i=1}^\infty \mathcal{H}^s_\delta(E_i)$ which shows $\mathcal{H}^s_\delta$ is an outer measure. Now notice that $\mathcal{H}^s_\delta(E)$ is increasing as $\delta \to 0$. Picking a sequence δ_k decreasing to 0, the monotone convergence theorem implies $\mathcal{H}^s(E) \leq \sum_{i=1}^\infty \mathcal{H}^s(E_i)$. ■

The outer measure $\mathcal{H}^s$ is called s dimensional Hausdorff measure when restricted to the σ algebra of $\mathcal{H}^s$ measurable sets. It is automatically a complete measure meaning that if $E \subseteq F$ where $\mathcal{H}^s(F) = 0$ then E is measurable. This follows from Theorem 9.8.4.

Next I will show the σ algebra of $\mathcal{H}^s$ measurable sets includes the Borel sets.

16.7 Properties of Hausdorff Measure

Using Theorem 9.9.1 on Page 262, the following is obtained.

Theorem 16.7.1 *The σ algebra of $\mathcal{H}^s$ measurable sets contains the Borel sets and $\mathcal{H}^s$ has the property that for all $E \subseteq \mathbb{R}^p$, there exists a Borel set $F \supseteq E$ such that $\mathcal{H}^s(F) = \mathcal{H}^s(E)$.*

Proof: Let $\operatorname{dist}(A,B) = 2\delta_0 > 0$. Is it the case that $\mathcal{H}^s(A) + \mathcal{H}^s(B) = \mathcal{H}^s(A \cup B)$? This is what is needed to use Theorem 9.9.1 about measurable sets including the Borel sets.

Let $\{C_j\}_{j=1}^\infty$ be a covering of $A \cup B$ such that $\operatorname{diam}(C_j) \leq \delta < \delta_0$ for each j and

$$\mathcal{H}^s_\delta(A \cup B) + \varepsilon > \sum_{j=1}^\infty \beta(s)(r(C_j))^s.$$

Thus $\mathcal{H}^s_\delta(A \cup B) + \varepsilon > \sum_{j \in J_1} \beta(s)(r(C_j))^s + \sum_{j \in J_2} \beta(s)(r(C_j))^s$ where

$$J_1 = \{j : C_j \cap A \neq \emptyset\},\ J_2 = \{j : C_j \cap B \neq \emptyset\}.$$

Recall $\operatorname{dist}(A,B) = 2\delta_0$, $J_1 \cap J_2 = \emptyset$. It follows $\mathcal{H}^s_\delta(A \cup B) + \varepsilon > \mathcal{H}^s_\delta(A) + \mathcal{H}^s_\delta(B)$. Letting $\delta \to 0$, and noting $\varepsilon > 0$ was arbitrary, yields

$$\mathcal{H}^s(A \cup B) \geq \mathcal{H}^s(A) + \mathcal{H}^s(B).$$

Equality holds because $\mathscr{H}^s$ is an outer measure. By Theorem 9.9.1, $\mathscr{H}^s$ is a Borel measure.

To verify the second assertion, note there is no loss of generality in letting $\mathscr{H}^s(E) < \infty$. Let $E \subseteq \cup_{j=1}^{\infty} C_j$, $r(C_j) < \delta$, and

$$\mathscr{H}^s_\delta(E) + \delta > \sum_{j=1}^{\infty} \beta(s)(r(C_j))^s. \tag{16.7}$$

Let $F_\delta = \cup_{j=1}^{\infty} \overline{C_j}$. Thus $F_\delta \supseteq E$ and

$$\mathscr{H}^s_\delta(E) \leq \mathscr{H}^s_\delta(F_\delta) \leq \sum_{j=1}^{\infty} \beta(s)(r(\overline{C_j}))^s = \sum_{j=1}^{\infty} \beta(s)(r(C_j))^s < \delta + \mathscr{H}^s_\delta(E).$$

Let $\delta_k \to 0$ and let $F = \cap_{k=1}^{\infty} F_{\delta_k}$. Then $F \supseteq E$ and

$$\mathscr{H}^s_{\delta_k}(E) \leq \mathscr{H}^s_{\delta_k}(F) \leq \mathscr{H}^s_{\delta_k}(F_\delta) \leq \delta_k + \mathscr{H}^s_{\delta_k}(E).$$

Letting $k \to \infty$, $\mathscr{H}^s(E) \leq \mathscr{H}^s(F) \leq \mathscr{H}^s(E)$

We can also arrange to have F containing E be a G_δ set. In 16.7, replace C_j with $C_j + B(\mathbf{0}, \eta_j)$ which is an open set having diameter no more than $\operatorname{diam}(C_j) + 2\eta_j$ so by taking η_j small enough, we can replace each C_j with an open set O_j in such a way as to preserve 16.7 with C_j replaced with O_j and also $r(O_j) < \delta$. Then letting $V_\delta \equiv \cup_j O_j$,

$$\mathscr{H}^s_\delta(E) \leq \mathscr{H}^s_\delta(V_\delta) \leq \sum_{j=1}^{\infty} \beta(s)(r(O_j))^s < \delta + \mathscr{H}^s_\delta(E).$$

Then let $G = \cap_k V_{\delta_k}$ where $\delta_k \to 0$ and let the V_{δ_k} be decreasing as k increases, each V_δ containing E. Then for each δ,

$$\mathscr{H}^s_{\delta_k}(E) \leq \mathscr{H}^s_{\delta_k}(G) < \delta + \mathscr{H}^s_{\delta_k}(E)$$

Let $k \to \infty$ to find $\mathscr{H}^s(E) \leq \mathscr{H}^s(G) \leq \mathscr{H}^s(E)$ as before. ■

A measure satisfying the conclusion of Theorem 16.7.1 is called a Borel regular measure.

16.8 $\mathscr{H}^p$ and m_p

Next I will compare $\mathscr{H}^p$ and m_p. First recall this covering theorem which is a summary of Corollary 9.12.5 found on Page 270.

Theorem 16.8.1 *Let $E \subseteq \mathbb{R}^p$ and let $\mathscr{F}$ be a collection of balls of bounded radii such that $\mathscr{F}$ covers E in the sense of Vitali. Then there exists a countable collection of disjoint balls from $\mathscr{F}$, $\{B_j\}_{j=1}^{\infty}$, such that $\overline{m_p}(E \setminus \cup_{j=1}^{\infty} B_j) = 0$.*

Lemma 16.8.2 *Every open set U in $\mathbb{R}^p$ is a countable disjoint union of half open boxes of the form $Q \equiv \prod_{i=1}^{p} [a_i, a_i + 2^{-k})$ where $a_i = l2^{-k}$ for l some integer.*

Proof: It is clear that there exists $\mathscr{Q}_k$ a countable disjoint collection of these half open boxes each of sides of length 2^{-k} whose union is all of $\mathbb{R}^p$. Let $\mathscr{B}_1$ be those sets of $\mathscr{Q}_1$ which are contained in U, if any. Having chosen $\mathscr{B}_{k-1}$, let $\mathscr{B}_k$ consist of those sets of $\mathscr{Q}_k$ which are contained in U such that none of these are contained in $\mathscr{B}_{k-1}$. Then $\cup_{k=1}^{\infty} \mathscr{B}_k$ is a countable collection of disjoint boxes of the right sort whose union is U. This is because if R is a box of $\mathscr{Q}_k$ and $\hat{R}$ is a box of $\mathscr{Q}_{k-1}$, then either $R \subseteq \hat{R}$ or $R \cap \hat{R} = \emptyset$. ■

Lemma 16.8.3 *If $S \subseteq \mathbb{R}^p$ and $m_p(S) = 0$, then $\mathcal{H}^p(S) = \mathcal{H}^p_\delta(S) = 0$. Also, there exists a constant k such that $\mathcal{H}^p(E) \leq k m_p(E)$ for all E Borel $k \equiv \frac{\beta(p)}{\alpha(p)}$. Also, if $Q_0 \equiv [0,1)^p$, the unit cube, then $\infty > \mathcal{H}^p([0,1)^p) > 0$.*

Proof: Suppose first $m_p(S) = 0$. Without loss of generality, S is bounded. Then by outer regularity, there exists a bounded open V containing S and $m_p(V) < \varepsilon$. For each $\boldsymbol{x} \in S$, there exists a ball $B_{\boldsymbol{x}}$ such that $\widehat{B_{\boldsymbol{x}}} \subseteq V$ and $\delta > r\left(\widehat{B_{\boldsymbol{x}}}\right)$. By the Vitali covering theorem there is a sequence of disjoint balls $\{B_k\}$ such that $\left\{\widehat{B_k}\right\}$ covers S. Here $\widehat{B_k}$ has the same center as B_k but 5 times the radius. Then letting $\alpha(p)$ be the Lebesgue measure of the unit ball in $\mathbb{R}^p$

$$\begin{aligned}\mathcal{H}^p_\delta(S) &\leq \sum_k \beta(p)\, r\left(\widehat{B_k}\right)^p = \frac{\beta(p)}{\alpha(p)} 5^p \sum_k \alpha(p)\, r(B_k)^p \\ &\leq \frac{\beta(p)}{\alpha(p)} 5^p m_p(V) < \frac{\beta(p)}{\alpha(p)} 5^p \varepsilon\end{aligned}$$

Since ε is arbitrary, this shows $\mathcal{H}^p_\delta(S) = 0$ and now it follows that

$$\mathcal{H}^p(S) \equiv \lim_{\delta \to 0} \mathcal{H}^p_\delta(S) = 0.$$

Letting U be an open set and $\delta > 0$, consider all balls B contained in U which have diameters less than δ. This is a Vitali covering of U and therefore by Theorem 16.8.1, there exists $\{B_i\}$, a sequence of disjoint balls of radii less than δ contained in U such that $\cup_{i=1}^\infty B_i$ differs from U by a set of Lebesgue measure zero. Let $\alpha(p)$ be the Lebesgue measure of the unit ball in $\mathbb{R}^p$. Then from what was just shown,

$$\begin{aligned}\mathcal{H}^p_\delta(U) &= \mathcal{H}^p_\delta(\cup_i B_i) \leq \sum_{i=1}^\infty \beta(p)\, r(B_i)^p = \frac{\beta(p)}{\alpha(p)} \sum_{i=1}^\infty \alpha(p)\, r(B_i)^p \\ &= \frac{\beta(p)}{\alpha(p)} \sum_{i=1}^\infty m_p(B_i) = \frac{\beta(p)}{\alpha(p)} m_p(U) \equiv k m_p(U),\ k \equiv \frac{\beta(p)}{\alpha(p)}\end{aligned}$$

Now letting E be Lebesgue measurable, it follows from the outer regularity of m_p there exists a decreasing sequence of open sets, $\{V_i\}$ containing E such such that $m_p(V_i) \to m_p(E)$. Then from the above,

$$\mathcal{H}^p_\delta(E) \leq \lim_{i \to \infty} \mathcal{H}^p_\delta(V_i) \leq \lim_{i \to \infty} k m_p(V_i) = k m_p(E).$$

Since $\delta > 0$ is arbitrary, it follows that also $\mathcal{H}^p(E) \leq k m_p(E)$. This proves the first part of the lemma and that $\infty > \mathcal{H}^p([0,1)^p)$.

If $\mathcal{H}^p([0,1)^p) = 0$, it follows $\mathcal{H}^p(\mathbb{R}^p) = 0$ because $\mathbb{R}^p$ is the countable union of translates of $Q_0 \equiv [0,1)^p$ and it is clear that $\mathcal{H}^p$ is translation invariant. Since each $\mathcal{H}^p_\delta$ is no larger than $\mathcal{H}^p$, $\mathcal{H}^p_\delta(\mathbb{R}^p) = 0$. Therefore, there exists a sequence of sets, $\{C_i\}$ each having diameter less than δ such that the union of these sets equals $\mathbb{R}^p$ but $1 > \sum_{i=1}^\infty \beta(p)\, r(C_i)^p$. Now let B_i be a ball having radius r_i equal to $\operatorname{diam}(C_i) = 2r(C_i)$ which contains C_i. These B_i cover $\mathbb{R}^p$, $\frac{1}{2} r_i = r(C_i)$. It follows that

$$1 > \sum_{i=1}^\infty \beta(p)\, r(C_i)^p = \sum_{i=1}^\infty \frac{\beta(p)}{\alpha(p)\, 2^p} m_p(B_i) = \infty,$$

a contradiction. This shows that $\mathscr{H}^p([0,1)^p) > 0$. ■

Note that the above shows that $\mathscr{H}^p([-n,n)^p)$ is always a finite positive real number for $n \in \mathbb{N}$.

Theorem 16.8.4 *By choosing $\beta(p)$ properly, one can obtain $\mathscr{H}^p = m_p$ on all Lebesgue measurable sets.*

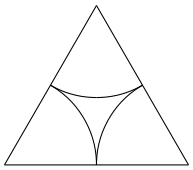
Proof: Define $l = \frac{m_p(Q_0)}{\mathscr{H}^p(Q_0)}$ where $Q_0 = [0,1)^p$ is the half open unit cube in $\mathbb{R}^p$. It follows then that $l = \frac{m_p(Q)}{\mathscr{H}^p(Q)}$ where $Q = \prod_{i=1}^p [a_i, a_i + 2^{-k})$ where $a_i = l2^{-k}$ for l some integer because of translation invariance of both measures and that Q_0 is the union of such Q. By Lemma 16.8.2, $l\mathscr{H}^p(V) = m_p(V)$ for any V open. Letting V_n be an increasing sequence of bounded open sets whose union is $\mathbb{R}^p$, it follows that the set of Borel E satisfying $l\mathscr{H}^p(E \cap V_n) = m_p(E \cap V_n)$ is a σ algebra which contains the open sets and so this equation is true for all Borel sets. Letting $n \to \infty, l\mathscr{H}^p(F) = m_p(F)$ for any Borel F. For E Lebesgue measurable, there is F Borel contained in E with $m_p(E \setminus F) = 0$ and so F is $\mathscr{H}^p$ measurable as is $E \setminus F$ because $\mathscr{H}^p(E \setminus F) = 0$ from Lemma 16.8.3 and $\mathscr{H}^p$ is an outer measure. Recall that sets having outer measure 0 end up being measurable sets. Thus $E = F \cup (E \setminus F)$ is $\mathscr{H}^p$ measurable also. This implies from Theorem 16.7.1 and Proposition 11.1.2 that if E is Lebesgue measurable, there is Borel $F \supseteq E$ such that $m_p(E) = m_p(F) = l\mathscr{H}^p(F) = l\mathscr{H}^p(E)$. Now choose $\beta(p)$ to make the constant $l = 1$. ■

The exact determination of $\beta(p)$ is more technical.

16.9 Technical Considerations

Let $\alpha(p)$ be the volume of the unit ball in $\mathbb{R}^p$. Thus the volume of $B(\mathbf{0}, r)$ in $\mathbb{R}^p$ is $\alpha(p)r^p$ from the change of variables formula. There is a very important and interesting inequality known as the isodiametric inequality which says that if A is any set in $\mathbb{R}^p$, then

$$\overline{m}_p(A) \leq \alpha(p)(2^{-1}\mathrm{diam}(A))^p = \alpha(p)\, r(A)^p.$$



This inequality may seem obvious at first but it is not really. The reason it is not is that there are sets which are not subsets of any sphere having the same diameter as the set. For example, consider an equilateral triangle. You have to include the vertices and so the center of such a ball would need to be closer to each vertex than the radius of the small circles. See the above picture

Lemma 16.9.1 *Let $f : \mathbb{R}^{p-1} \to [0,\infty)$ be Borel measurable and let*

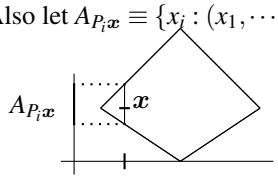
$$S = \{(\boldsymbol{x}, y) : |y| < f(\boldsymbol{x})\}.$$

Then S is a Borel set in $\mathbb{R}^p$.

Proof: Set s_k be an increasing sequence of Borel measurable functions converging pointwise to f. $s_k(\boldsymbol{x}) = \sum_{m=1}^{N_k} c_m^k \mathscr{X}_{E_m^k}(\boldsymbol{x})$. Let $S_k = \cup_{m=1}^{N_k} E_m^k \times (-c_m^k, c_m^k)$. Then $(\boldsymbol{x}, y) \in S_k$

if and only if $f(\boldsymbol{x}) > 0$ and $|y| < s_k(\boldsymbol{x}) \leq f(\boldsymbol{x})$. It follows that $S_k \subseteq S_{k+1}$ and $S = \cup_{k=1}^{\infty} S_k$. But each S_k is a Borel set and so S is also a Borel set. ■

Let P_i be the projection onto span $(\boldsymbol{e}_1, \cdots, \boldsymbol{e}_{i-1}, \boldsymbol{e}_{i+1}, \cdots, \boldsymbol{e}_p)$ where the $\boldsymbol{e}_k$ are the standard basis vectors in $\mathbb{R}^p$, $\boldsymbol{e}_k$ being the vector having a 1 in the k^{th} slot and a 0 elsewhere. Thus $P_i\boldsymbol{x} \equiv \sum_{j \neq i} x_j \boldsymbol{e}_j$. Also let $A_{P_i\boldsymbol{x}} \equiv \{x_i : (x_1, \cdots, x_i, \cdots, x_p) \in A\}$



$P_i\boldsymbol{x} \in$ span$\{\boldsymbol{e}_1, \cdots, \boldsymbol{e}_{i-1}\boldsymbol{e}_{i+1}, \cdots, \boldsymbol{e}_n\}$.

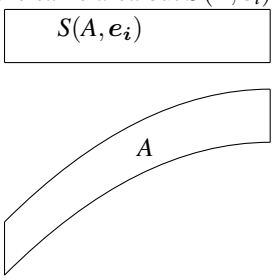
Lemma 16.9.2 *Let $A \subseteq \mathbb{R}^p$ be a Borel set. Then $P_i\boldsymbol{x} \to m(A_{P_i\boldsymbol{x}})$ is a Borel measurable function defined on $P_i(\mathbb{R}^p)$.*

Proof: From Proposition 10.14.3, $A_{P_i\boldsymbol{x}}$ is measurable if S is product measurable. By Theorem 10.14.8, the Borel sets are product measurable and $P_i\boldsymbol{x} \to m(A_{P_i\boldsymbol{x}})$ is measurable, in fact product measurable. Therefore, the desired conclusion of this lemma follows. ■

16.9.1 Steiner Symmetrization

Definition 16.9.3 *Define $S(A, \boldsymbol{e}_i) \equiv \{\boldsymbol{x} \equiv P_i\boldsymbol{x} + y\boldsymbol{e}_i : |y| < 2^{-1} m(A_{P_i\boldsymbol{x}})\}$.*

Here is a picture of the idea used in producing $S(A, \boldsymbol{e}_i)$ from A. The one on the top is $S(A, \boldsymbol{e}_i)$. The two sets have the same area but $S(A, \boldsymbol{e}_i)$ smaller diameter than A.



Lemma 16.9.4 *Let A be a Borel subset of $\mathbb{R}^p$. Then $S(A, \boldsymbol{e}_i)$ satisfies*

$$P_i\boldsymbol{x} + y\boldsymbol{e}_i \in S(A, \boldsymbol{e}_i) \text{ if and only if } P_i\boldsymbol{x} - y\boldsymbol{e}_i \in S(A, \boldsymbol{e}_i),$$

$$S(A, \boldsymbol{e}_i) \text{ is a Borel set in } \mathbb{R}^p,$$

$$m_p(S(A, \boldsymbol{e}_i)) = m_p(A), \tag{16.8}$$

$$diam(S(A, \boldsymbol{e}_i)) \leq diam(A). \tag{16.9}$$

Proof: The first assertion is obvious from the definition. The Borel measurability of $S(A, \boldsymbol{e}_i)$ follows from the definition and Lemmas 16.9.2 and 16.9.1. To show 16.8,

$$\begin{aligned} m_p(S(A, \boldsymbol{e}_i)) &= \int_{P_i\mathbb{R}^p} \int_{-2^{-1}m(A_{P_i\boldsymbol{x}})}^{2^{-1}m(A_{P_i\boldsymbol{x}})} dx_i dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_p \\ &= \int_{P_i\mathbb{R}^p} m(A_{P_i\boldsymbol{x}}) dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_p \\ &= \int_{P_i\mathbb{R}^p} \int_{\mathbb{R}} \mathcal{X}_A dx_i dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_p = m_p(A) \end{aligned}$$

Now suppose $\boldsymbol{x}_1$ and $\boldsymbol{x}_2 \in S(A,\boldsymbol{e}_i)$

$$\boldsymbol{x}_1 = P_i\boldsymbol{x}_1 + y_1\boldsymbol{e}_i,\ \boldsymbol{x}_2 = P_i\boldsymbol{x}_2 + y_2\boldsymbol{e}_i.$$

Now $y_1 \in \left[-\frac{m(A_{P_i\boldsymbol{x}_1})}{2}, \frac{m(A_{P_i\boldsymbol{x}_1})}{2}\right], y_2 \in \left[-\frac{m(A_{P_i\boldsymbol{x}_2})}{2}, \frac{m(A_{P_i\boldsymbol{x}_2})}{2}\right]$. There exists

$$x_{1i} \in [\inf A_{P_i\boldsymbol{x}_1}, \sup A_{P_i\boldsymbol{x}_1}] \text{ and } x_{2i} \in [\inf A_{P_i\boldsymbol{x}_2}, \sup A_{P_i\boldsymbol{x}_2}]$$

such that $x_{1i} \in A_{P_i\boldsymbol{x}_1}$ and $x_{2i} \in A_{P_i\boldsymbol{x}_2}$. The second pair of intervals is at least as long as the corresponding interval in the first pair and the second pair are not necessarily centered at the same point. Therefore, such an x_{1i} and x_{2i} can be chosen such that $|x_{2i} - x_{1i}| \geq |y_1 - y_2|$ and so $\hat{\boldsymbol{x}}_1 \equiv P_i\boldsymbol{x}_1 + x_{1i}\boldsymbol{e}_i$ and $\hat{\boldsymbol{x}}_2 \equiv P_i\boldsymbol{x}_2 + x_{2i}\boldsymbol{e}_i$ are in A and $|\hat{\boldsymbol{x}}_1 - \hat{\boldsymbol{x}}_2| \geq |\boldsymbol{x}_1 - \boldsymbol{x}_2|$ so the diameter of $S(A,\boldsymbol{e}_i)$ is no more than the diameter of A as claimed. ■

The next lemma says that if A is already symmetric with respect to the j^{th} direction, then this symmetry is not destroyed by taking $S(A,\boldsymbol{e}_i)$.

Lemma 16.9.5 *Suppose A is a Borel set in $\mathbb{R}^p$ such that $P_j\boldsymbol{x} + \boldsymbol{e}_j x_j \in A$ if and only if $P_j\boldsymbol{x} + (-x_j)\boldsymbol{e}_j \in A$. Then if $i \neq j$, $P_j\boldsymbol{x} + \boldsymbol{e}_j x_j \in S(A,\boldsymbol{e}_i)$ if and only if $P_j\boldsymbol{x} + (-x_j)\boldsymbol{e}_j \in S(A,\boldsymbol{e}_i)$.*

Proof: By definition, $P_j\boldsymbol{x} + \boldsymbol{e}_j x_j \in S(A,\boldsymbol{e}_i)$ if and only if $|x_i| < 2^{-1}m(A_{P_i(P_j\boldsymbol{x}+\boldsymbol{e}_j x_j)})$. Now $x_i \in A_{P_i(P_j\boldsymbol{x}+\boldsymbol{e}_j x_j)}$ if and only if $x_i \in A_{P_i(P_j\boldsymbol{x}+(-x_j\boldsymbol{e}_j))}$ by the assumption on A which says that A is symmetric in the $\boldsymbol{e}_j$ direction. Hence $P_j\boldsymbol{x} + \boldsymbol{e}_j x_j \in S(A,\boldsymbol{e}_i)$ if and only if $|x_i| < 2^{-1}m(A_{P_i(P_j\boldsymbol{x}+(-x_j)\boldsymbol{e}_j)})$ if and only if $P_j\boldsymbol{x} + (-x_j)\boldsymbol{e}_j \in S(A,\boldsymbol{e}_i)$. ■

16.9.2 The Isodiametric Inequality

The next theorem is called the isodiametric inequality. It is the key result used to compare Lebesgue and Hausdorff measures.

Theorem 16.9.6 *Let A be any Lebesgue measurable set in $\mathbb{R}^p$. Then*

$$m_p(A) \leq \alpha(p)(r(A))^p.$$

Proof: Suppose first that A is Borel. Let $A_1 = S(A,\boldsymbol{e}_1)$ and $A_k = S(A_{k-1},\boldsymbol{e}_k)$. Then by Lemma 16.9.4, A_p is a Borel set, $\operatorname{diam}(A_p) \leq \operatorname{diam}(A)$, $m_p(A_p) = m_p(A)$ and A_p is symmetric. Thus $\boldsymbol{x} \in A_p$ if and only if $-\boldsymbol{x} \in A_p$. It follows that $A_p \subseteq \overline{B(\boldsymbol{0}, r(A_p))}$. If $\boldsymbol{x} \in A_p \setminus \overline{B(\boldsymbol{0}, r(A_p))}$, then $-\boldsymbol{x} \in A_p \setminus \overline{B(\boldsymbol{0}, r(A_p))}$ and so $\operatorname{diam}(A_p) \geq 2|\boldsymbol{x}| > \operatorname{diam}(A_p)$. Therefore, there is no such $\boldsymbol{x}$ and $m_p(A_p) \leq \alpha(p)(r(A_p))^p \leq \alpha(p)(r(A))^p$. It remains to establish this inequality for arbitrary measurable sets. Letting A be such a set, let $\{K_k\}$ be an increasing sequence of compact subsets of A such that $m_p(A) = \lim_{k\to\infty} m_p(K_k)$. Then

$$m_p(A) = \lim_{k\to\infty} m_p(K_k) \leq \limsup_{k\to\infty} \alpha(p)(r(K_k))^p \leq \alpha(p)(r(A))^p.$$ ■

16.10 The Proper Value of $\beta(p)$

I will show that the proper determination of $\beta(p)$ is $\alpha(p)$, the volume of the unit ball. Since $\beta(p)$ has been adjusted such that $l = 1$ in Theorem 16.8.4, $m_p(B(\boldsymbol{0},1)) = \mathscr{H}^p(B(\boldsymbol{0},1))$. There exists a covering of $B(\boldsymbol{0},1)$ of sets of radii less than δ, $\{C_i\}_{i=1}^{\infty}$ such that

$$\mathscr{H}^p_\delta(B(\boldsymbol{0},1)) + \varepsilon > \sum_i \beta(p)\, r(C_i)^p$$

Then by Theorem 16.9.6, the isodiametric inequality,

$$\begin{aligned}\mathscr{H}_\delta^p(B(\mathbf{0},1))+\varepsilon &> \sum_i \beta(p)\, r(C_i)^p = \frac{\beta(p)}{\alpha(p)}\sum_i \alpha(p)\, r(\overline{C}_i)^p \\ &\geq \frac{\beta(p)}{\alpha(p)}\sum_i m_p(\overline{C}_i) \geq \frac{\beta(p)}{\alpha(p)} m_p(B(\mathbf{0},1)) = \frac{\beta(p)}{\alpha(p)}\mathscr{H}^p(B(\mathbf{0},1))\end{aligned}$$

Now taking the limit as $\delta \to 0$, $\mathscr{H}^p(B(\mathbf{0},1))+\varepsilon \geq \frac{\beta(p)}{\alpha(p)}\mathscr{H}^p(B(\mathbf{0},1))$ and since $\varepsilon > 0$ is arbitrary, this shows $\alpha(p) \geq \beta(p)$.

By the Vitali covering theorem in Corollary 9.12.5, there exists a sequence of disjoint balls of radius no more than δ, $\{B_i\}$ such that $B(\mathbf{0},1) = (\cup_{i=1}^\infty B_i) \cup N$. where $m_p(N) = 0$. Then $\mathscr{H}_\delta^p(N) = 0$ can be concluded because $\mathscr{H}_\delta^p \leq \mathscr{H}^p$ and Lemma 16.8.3. Using $m_p(B(\mathbf{0},1)) = \mathscr{H}^p(B(\mathbf{0},1))$ again,

$$\begin{aligned}\mathscr{H}_\delta^p(B(\mathbf{0},1)) &= \mathscr{H}_\delta^p(\cup_i B_i) \leq \sum_{i=1}^\infty \beta(p)\, r(B_i)^p \\ &= \frac{\beta(p)}{\alpha(p)}\sum_{i=1}^\infty \alpha(p)\, r(B_i)^p = \frac{\beta(p)}{\alpha(p)}\sum_{i=1}^\infty m_p(B_i) \\ &= \frac{\beta(p)}{\alpha(p)} m_p(\cup_i B_i) = \frac{\beta(p)}{\alpha(p)} m_p(B(\mathbf{0},1)) = \frac{\beta(p)}{\alpha(p)}\mathscr{H}^p(B(\mathbf{0},1))\end{aligned}$$

which implies $\alpha(p) \leq \beta(p)$ and so the two are equal. This proves that if $\alpha(p) = \beta(p)$, then the $\mathscr{H}^p = m_p$ on the measurable sets of $\mathbb{R}^p$.

This gives another way to think of Lebesgue measure which is a particularly nice way because it is coordinate free, depending only on the notion of distance.

For $s < p$, note that $\mathscr{H}^s$ is not a Radon measure because it will not generally be finite on compact sets. For example, let $p = 2$ and consider $\mathscr{H}^1(L)$ where L is a line segment joining $(0,0)$ to $(1,0)$. Then $\mathscr{H}^1(L)$ is no smaller than $\mathscr{H}^1(L)$ when L is considered a subset of $\mathbb{R}^1, p = 1$. Thus by what was just shown, $\mathscr{H}^1(L) \geq 1$. Hence $\mathscr{H}^1([0,1]\times[0,1]) = \infty$. The situation is this: L is a one-dimensional object inside $\mathbb{R}^2$ and $\mathscr{H}^1$ is giving a one-dimensional measure of this object. In fact, Hausdorff measures can make such heuristic remarks as these precise. Define the Hausdorff dimension of a set, A, as $\dim(A) = \inf\{s : \mathscr{H}^s(A) = 0\}$.

16.11 A Formula for $\alpha(p)$

What is $\alpha(p)$ for p a positive integer? Let p be a positive integer. Theorem 14.5.1 on Page 411 says that

Theorem 16.11.1 *$\alpha(p) = \pi^{p/2}(\Gamma(p/2+1))^{-1}$ where $\Gamma(s)$ is the gamma function $\Gamma(s) = \int_0^\infty e^{-t}t^{s-1}dt$.*

From now on, in the definition of Hausdorff measure, it will always be the case that $\beta(s) = \alpha(s)$. As shown above, this is the right thing to have $\beta(s)$ to equal if s is a positive integer because this yields the important result that Hausdorff measure is the same as Lebesgue measure. Note the formula, $\pi^{s/2}(\Gamma(s/2+1))^{-1}$ makes sense for any $s \geq 0$.

Chapter 17

The Area Formula

I am grateful to those who have found errors in this material, some of which were egregious. I would not have found these mistakes because I never teach this material and I don't use it in my research. I do think it is wonderful mathematics however.

17.1 Estimates for Hausdorff Measure

This section is on estimates which relate Hausdorff measure to Lebesgue measure. This will allow a geometric motivation for measures on Lipschitz manifolds.

The main case will be for $\boldsymbol{h}$ a Lipschitz function, $|\boldsymbol{h}(\boldsymbol{x})-\boldsymbol{h}(\boldsymbol{y})| \leq K|\boldsymbol{x}-\boldsymbol{y}|$ defined on $\mathbb{R}^n$. This is no loss of generality because of Theorem 16.3.5. However, the main presentation will include more general situations than this. One uses the differentiability of $\boldsymbol{h}$ off a set of measure zero to show the existence of disjoint Borel sets E on which $\boldsymbol{h}$ is Lipschitz with its inverse also being Lipschitz on $\boldsymbol{h}(E)$.

The following lemma states that Lipschitz maps take sets of measure zero to sets of measure zero. It also gives a convenient estimate. This involves the consideration of $\mathscr{H}^n$ as an outer measure. Thus it is not necessary to know the set B is measurable.

In fact, one only needs to have $\boldsymbol{h}$ locally Lipschitz in much of what follows.

Definition 17.1.1 *Let $\boldsymbol{h}:\mathbb{R}^n \to \mathbb{R}^m$. This function is said to be locally Lipschitz if for every $\boldsymbol{x} \in \mathbb{R}^n$, there exists a ball $B_{\boldsymbol{x}}$ containing $\boldsymbol{x}$ and a constant $K_{\boldsymbol{x}}$ such that for all $\boldsymbol{y},\boldsymbol{z} \in B_{\boldsymbol{x}}$,*

$$|\boldsymbol{h}(\boldsymbol{z})-\boldsymbol{h}(\boldsymbol{y})| \leq K_{\boldsymbol{x}}|\boldsymbol{z}-\boldsymbol{y}|$$

Lemma 17.1.2 *If $\boldsymbol{h}$ is Lipschitz with Lipschitz constant K then for $B \subseteq \mathbb{R}^n$,*

$$\mathscr{H}^n(\boldsymbol{h}(B)) \leq K^n \mathscr{H}^n(B)$$

Also, if T is a set in $\mathbb{R}^n$, $m_n(T)=0$, then $\mathscr{H}^n(\boldsymbol{h}(T))=0$. It is not necessary that $\boldsymbol{h}$ be one to one.

Proof: If $\mathscr{H}^n(B)=\infty$, there is nothing to show. Assume $\mathscr{H}^n(B)<\infty$. Let $\{C_i\}_{i=1}^{\infty}$ cover B with each having diameter less than δ and let this cover be such that

$$\sum_i \beta(n)\frac{1}{2}\operatorname{diam}(C_i)^n < \mathscr{H}^n_\delta(B)+\varepsilon$$

Then $\{\boldsymbol{h}(C_i)\}$ covers $\boldsymbol{h}(B)$ and each set has diameter no more than $K\delta$. Then

$$\begin{aligned}\mathscr{H}^n_{K\delta}(\boldsymbol{h}(B)) &\leq \sum_i \beta(n)\left(\frac{1}{2}\operatorname{diam}(\boldsymbol{h}(C_i))\right)^n \\ &\leq K^n\sum_i \beta(n)\left(\frac{1}{2}\operatorname{diam}(C_i)\right)^n \leq K^n\left(\mathscr{H}^n_\delta(B)+\varepsilon\right)\end{aligned}$$

Since ε is arbitrary, this shows that $\mathscr{H}^n_{K\delta}(\boldsymbol{h}(B)) \leq K^n\mathscr{H}^n_\delta(B)$. Now take a limit as $\delta \to 0$. The second claim follows from $m_n=\mathscr{H}^n$ on Lebesgue measurable sets of $\mathbb{R}^n$. ■

Lemma 17.1.3 *If $\boldsymbol{h}$ is locally Lipschitz and $m_n(T)=0$, then $\mathscr{H}^n(\boldsymbol{h}(T))=0$. It is not necessary that $\boldsymbol{h}$ be one to one.*

Proof: Let $T_k \equiv \{\boldsymbol{x} \in T : \boldsymbol{h} \text{ has Lipschitz constant } k \text{ near } \boldsymbol{x}\}$.Thus $T = \cup_k T_k$. I will show $\boldsymbol{h}(T_k)$ has $\mathscr{H}^n$ measure zero and then it will follow that $\boldsymbol{h}(T) = \cup_{k=1}^{\infty} \boldsymbol{h}(T_k)$, the $\boldsymbol{h}(T_k)$ increasing in k, must also have measure zero.

Let $\varepsilon > 0$ be given. By outer regularity, there exists an open set V containing T_k such that $m_n(V) < \varepsilon$. For $\boldsymbol{x} \in T_k$ it follows there exists $r_{\boldsymbol{x}} < 1$ such that the ball centered at $\boldsymbol{x}$ with radius $r_{\boldsymbol{x}}$ is contained in V and in this ball, $\boldsymbol{h}$ has Lipschitz constant k. By the Besicovitch covering theorem, Theorem 4.5.8, there are N_n sets of these balls $\{\mathscr{G}_1, \cdots, \mathscr{G}_{N_n}\}$ such that the balls in $\mathscr{G}_k$ are disjoint and the union of all balls in the N_n sets covers T_k. Then

$$\begin{aligned} \mathscr{H}^n(\boldsymbol{h}(T_k)) &\leq \sum_{k=1}^{N_n} \sum \{\mathscr{H}^n(\boldsymbol{h}(B)) : B \in \mathscr{G}_k\} \\ &\leq N_n k^n m_n(V) < N_n k^n \varepsilon \end{aligned}$$

Since ε is arbitrary, this shows that $\mathscr{H}^n(\boldsymbol{h}(T_k)) = 0$. Hence $\mathscr{H}^n(\boldsymbol{h}(T)) = 0$ also, since it is the limit of the $\mathscr{H}^n(\boldsymbol{h}(T_k))$. ■

Lemma 17.1.4 *If S is a Lebesgue measurable set in $\mathbb{R}^n$ and $\boldsymbol{h}$ is Lipschitz or locally Lipschitz then $\boldsymbol{h}(S)$ is $\mathscr{H}^n$ measurable. Also, if $\boldsymbol{h}$ is Lipschitz with constant K, $\mathscr{H}^n(\boldsymbol{h}(S)) \leq K^n m_n(S)$. It is not necessary that $\boldsymbol{h}$ be one to one.*

Proof: The estimate follows from Lemma 17.1.2 or 17.1.3 and the observation that, as shown before, Theorem 16.8.4, if S is Lebesgue measurable in $\mathbb{R}^n$, then $\mathscr{H}^n(S) = m_n(S)$. The estimate also shows that $\boldsymbol{h}$ maps sets of Lebesgue measure zero to sets of $\mathscr{H}^n$ measure zero. Why is $\boldsymbol{h}(S)$ $\mathscr{H}^n$ measurable if S is Lebesgue measurable? This follows from completeness of $\mathscr{H}^n$. Indeed, let F be F_σ and contained in S with $m_n(S \setminus F) = 0$. Then $\boldsymbol{h}(S) = \boldsymbol{h}(S \setminus F) \cup \boldsymbol{h}(F)$. The second set is Borel and the first has $\mathscr{H}^n$ measure zero. By completeness of $\mathscr{H}^n$, $\boldsymbol{h}(S)$ is $\mathscr{H}^n$ measurable. ■

Recall Theorem 1.5.5 on Page 23. This is stated here for convenience.

Theorem 17.1.5 *Let F be an $m \times p$ matrix where $m \geq p$. Then there exists an $m \times p$ matrix R and a $p \times p$ matrix U such that*

$$F = RU,\ U = U^*,$$

all eigenvalues of U are non negative, $U^2 = F^*F$, $R^*R = I$, and $|R\boldsymbol{x}| = |\boldsymbol{x}|$.

Thus, if $\boldsymbol{h} : \mathbb{R}^p \to \mathbb{R}^m, m \geq p$, and $D\boldsymbol{h}(\boldsymbol{x})$ exists, then $D\boldsymbol{h}(\boldsymbol{x}) = R(\boldsymbol{x})U(\boldsymbol{x})$ where

$$(U(\boldsymbol{x})\boldsymbol{u}, \boldsymbol{v}) = (U(\boldsymbol{x})\boldsymbol{v}, \boldsymbol{u}), (U(\boldsymbol{x})\boldsymbol{u}, \boldsymbol{u}) \geq 0$$

and $R^*R = I$ so R preserves lengths. Recall that R^* is the adjoint defined by $(R\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}, R^*\boldsymbol{y})$. This convention will be used in what follows.

Lemma 17.1.6 *In this situation where $R^*R = I$, $|R^*\boldsymbol{u}| \leq |\boldsymbol{u}|$.*

Proof: First note that $(\boldsymbol{u} - RR^*\boldsymbol{u}, RR^*\boldsymbol{u}) = (\boldsymbol{u}, RR^*\boldsymbol{u}) - |RR^*\boldsymbol{u}|^2 = |R^*\boldsymbol{u}|^2 - |R^*\boldsymbol{u}|^2 = 0$, and so

$$|\boldsymbol{u}|^2 = |\boldsymbol{u} - RR^*\boldsymbol{u} + RR^*\boldsymbol{u}|^2 = |\boldsymbol{u} - RR^*\boldsymbol{u}|^2 + |RR^*\boldsymbol{u}|^2 = |\boldsymbol{u} - RR^*\boldsymbol{u}|^2 + |R^*\boldsymbol{u}|^2. \blacksquare$$

Then the following corollary follows from Lemma 17.1.6.

Corollary 17.1.7 *Let $T \subseteq \mathbb{R}^m$. Then $\mathscr{H}^n(T) \geq \mathscr{H}^n(R^*T)$.*

Hausdorff measure makes possible a unified development of p dimensional area. As in the case of Lebesgue measure, the first step in this is to understand basic considerations related to linear transformations.

Lemma 17.1.8 *Let $R \in \mathscr{L}(\mathbb{R}^p, \mathbb{R}^m)$, $p \leq m$, and $R^*R = I$. Then if $A \subseteq \mathbb{R}^p$, $\mathscr{H}^p(RA) = \mathscr{H}^p(A)$. In fact, if $P : \mathbb{R}^p \to \mathbb{R}^m$ satisfies $|P\boldsymbol{x} - P\boldsymbol{y}| = |\boldsymbol{x} - \boldsymbol{y}|$, then $\mathscr{H}^p(PA) = \mathscr{H}^p(A)$.*

Proof: Now let P be an arbitrary mapping which preserves lengths, like R, and let A be bounded so $P(A)$ is also bounded. Then $P(A) \subseteq \cup_{j=1}^{\infty} C_j$, $r(C_j) < \delta$, and $\mathscr{H}^p_\delta(PA) + \varepsilon > \sum_{j=1}^{\infty} \alpha(p)(r(C_j))^p$.Since P preserves lengths, it follows P is one to one and P^{-1} is one to one on $P(\mathbb{R}^p)$ and P^{-1} also preserves lengths on $P(\mathbb{R}^p)$. Replacing each C_j with $C_j \cap (PA)$,

$$\mathscr{H}^p_\delta(PA) + \varepsilon > \sum_{j=1}^{\infty} \alpha(p) r(C_j \cap (PA))^p = \sum_{j=1}^{\infty} \alpha(p) r\left(P^{-1}(C_j \cap (PA))\right)^p \geq \mathscr{H}^p_\delta(A).$$

Thus $\mathscr{H}^p_\delta(PA) \geq \mathscr{H}^p_\delta(A)$. Similarly $\mathscr{H}^p_\delta(P^{-1}(PA)) \geq \mathscr{H}^p_\delta(PA)$ so $\mathscr{H}^p_\delta(A) \geq \mathscr{H}^p_\delta(PA)$. Letting $\delta \to 0$ yields the desired conclusion in the case where A is bounded. For the general case, let $A_r = A \cap B(0,r)$. Then $\mathscr{H}^p(PA_r) = \mathscr{H}^p(A_r)$. Now let $r \to \infty$. ■

Lemma 17.1.9 *Let $F \in \mathscr{L}(\mathbb{R}^p, \mathbb{R}^m), p \leq m$, and let $F = RU$ where R and U are described in Theorem 1.5.5 on Page 23. Then if $A \subseteq \mathbb{R}^p$ is Lebesgue measurable,*

$$\mathscr{H}^p(FA) = \det(U) m_p(A).$$

Proof: Using Theorem 11.13.2 on Page 350 and Theorem 16.8.4,

$$\mathscr{H}^p(FA) = \mathscr{H}^p(RUA) = \mathscr{H}^p(UA) = m_p(UA) = \det(U) m_p(A). \blacksquare$$

Definition 17.1.10 *Define J to equal $\det(U)$. Thus*

$$J = \det((F^*F)^{1/2}) = (\det(F^*F))^{1/2}.$$

This is the essential idea for the area formula, but in the area formula, we must consider $\boldsymbol{h} : \mathbb{R}^p \to \mathbb{R}^m$ for $\boldsymbol{h}$ nonlinear and so $\boldsymbol{h}(U)$ is not a subspace.

17.2 Comparison Theorems

First is a simple lemma which is fairly interesting which involves comparison of two linear transformations.

Lemma 17.2.1 *Suppose S, T are linear, defined on a finite dimensional normed linear space, S^{-1} exists, and let $\delta \in (0,1)$. Then whenever $\|S - T\|$ is small enough, it follows that*

$$\frac{|T\boldsymbol{v}|}{|S\boldsymbol{v}|} \in (1-\delta, 1+\delta) \tag{17.1}$$

for all $\boldsymbol{v} \neq \boldsymbol{0}$. Similarly if T^{-1} exists and $\|S - T\|$ is small enough,

$$\frac{|T\boldsymbol{v}|}{|S\boldsymbol{v}|} \in (1-\delta, 1+\delta)$$

Proof: Say S^{-1} exists. Then $v \to |Sv|$ is a norm. Then by equivalence of norms, Theorem 4.4.9, there exists $\eta > 0$ such that for all v, $|Sv| \geq \eta |v|$. Say $\|T-S\| < r < \delta\eta$

$$\frac{|Tv|}{|Sv|} = \frac{|Sv-(S-T)v|}{|Sv|} \geq \frac{|Sv|-\|T-S\||v|}{|Sv|} \geq \frac{|Sv|-\delta\eta|v|}{|Sv|} \geq \frac{|Sv|-\delta|Sv|}{|Sv|} = 1-\delta$$

$$\frac{|Tv|}{|Sv|} = \frac{|Sv+(T-S)v|}{|Sv|} \leq \frac{|Sv|+\|T-S\||v|}{|Sv|} \leq \frac{|Sv|+\delta\eta|v|}{|Sv|} \leq \frac{|Sv|+\delta|Sv|}{|Sv|} = 1+\delta$$

Next suppose that T^{-1} exists. Then, letting $\hat{\delta}$ be small enough, $\left(1-\hat{\delta}, 1+\hat{\delta}\right) \subseteq \left(\frac{1}{1+\delta}, \frac{1}{1-\delta}\right)$. From what was just shown, if $\|S-T\|$ is small enough,

$$\frac{|Sv|}{|Tv|} \in \left(1-\hat{\delta}, 1+\hat{\delta}\right) \subseteq \left(\frac{1}{1+\delta}, \frac{1}{1-\delta}\right) \text{ so } \frac{|Tv|}{|Sv|} \in (1-\delta, 1+\delta). \blacksquare$$

In short, the above lemma says that if one of S, T is invertible and the other is close to it, then it is also invertible and the quotient of $|Sv|$ and $|Tv|$ is close to 1. Then the following lemma is fairly obvious.

Lemma 17.2.2 *Let S, T be $n \times n$ matrices which are invertible. Then*

$$o(Tv) = o(Sv) = o(v)$$

and if L is a continuous linear transformation such that for $a < b$,

$$\sup_{v \neq 0} \frac{|Lv|}{|Sv|} < b, \ \inf_{v \neq 0} \frac{|Lv|}{|Sv|} > a$$

If $\|S-T\|$ is small enough, it follows that the same inequalities hold with S replaced with T. Here $\|\cdot\|$ denotes the operator norm.

Proof: Consider the first claim. For

$$\frac{|o(Tv)|}{|v|} = \frac{|o(Tv)|}{|Tv|}\frac{|Tv|}{|v|} \leq \frac{|o(Tv)|}{|Tv|}\|T\|$$

Thus $o(Tv) = o(v)$. It is similar for T replaced with S.

Consider the second claim. Pick δ sufficiently small. Then by Lemma 17.2.1

$$\sup_{v \neq 0} \frac{|Lv|}{|Tv|} = \sup_{v \neq 0} \frac{|Lv|}{|Sv|}\frac{|Sv|}{|Tv|} \leq (1+\delta) \sup_{v \neq 0} \frac{|Lv|}{|Sv|} < b$$

if δ is small enough. The other inequality is shown exactly similar. $\blacksquare$

17.3 A Decomposition

This follows [17] which is where I encountered this material. Let G be an open set and let $h : G \subseteq \mathbb{R}^n \to \mathbb{R}^m$ where $m \geq n$ be continuous. Let $Dh(x)$ exist for all $x \in A$ a Borel set contained in G. Also, $\mathcal{X}_A Dh(x)$, considered as a matrix has all of its entries Borel measurable if h is continuous because these are obtained as partial derivatives which are limits of continuous functions. Of course this is automatic if h is Lipschitz because then A

is all of G other than a set of measure zero. Recall that if $\boldsymbol{h}$ is Lipschitz continuous, $\boldsymbol{h}(E)$ is $\mathscr{H}^n$ Hausdorff measurable whenever E is n dimensional Lebesgue measurable.

For convenience $0 < \varepsilon < 1/4$ in what follows.

For $\boldsymbol{x} \in A$, let $D\boldsymbol{h}(\boldsymbol{x}) \equiv R(\boldsymbol{x}) U(\boldsymbol{x})$ where $R(\boldsymbol{x})$ preserves lengths and

$$U(\boldsymbol{x}) \equiv \left(D\boldsymbol{h}(\boldsymbol{x})^* D\boldsymbol{h}(\boldsymbol{x})\right)^{1/2}.$$

This is the right polar factorization of Theorem 1.5.5. Let A^+ denote those points of A for which $U(\boldsymbol{x})^{-1}$ exists, where $\det\left(D\boldsymbol{h}(\boldsymbol{x})^* D\boldsymbol{h}(\boldsymbol{x})\right) > 0$. Thus this is a Borel measurable subset of A. Note that $\det\left(D\boldsymbol{h}(\boldsymbol{x})^* D\boldsymbol{h}(\boldsymbol{x})\right) \geq 0$ for all $\boldsymbol{x} \in A$.

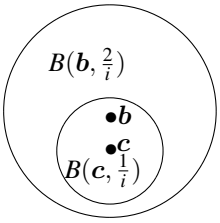
Let B be a Borel measurable subset of A^+ and let $\boldsymbol{b} \in B$. Let $\mathscr{S}$ be a countable dense subset of the space of symmetric invertible matrices and let $\mathscr{C}$ be a countable dense subset of B. The idea is to decompose B into countably many Borel sets E on which $\boldsymbol{h}$ is one to one and Lipschitz with $\boldsymbol{h}^{-1}$ Lipschitz on $\boldsymbol{h}(E)$. This will be done by establishing 17.6 given below where T is an invertible symmetric transformation. For $T \in \mathscr{S}$ and $\boldsymbol{c} \in \mathscr{C}$, define $E(T, \boldsymbol{c}, i)$ to be those $\boldsymbol{b} \in B\left(\boldsymbol{c}, \frac{1}{i}\right)$ such that for all $\boldsymbol{a} \in B\left(\boldsymbol{b}, \frac{2}{i}\right)$,

$$|\boldsymbol{h}(\boldsymbol{a}) - \boldsymbol{h}(\boldsymbol{b}) - D\boldsymbol{h}(\boldsymbol{b})(\boldsymbol{a} - \boldsymbol{b})| < \varepsilon |T(\boldsymbol{a} - \boldsymbol{b})| \tag{17.2}$$

and also $D\boldsymbol{h}(\boldsymbol{b})$ is close enough to T that the following hold.

$$\inf_{\boldsymbol{v} \neq \boldsymbol{0}} \frac{|D\boldsymbol{h}(\boldsymbol{b})\boldsymbol{v}|}{|T\boldsymbol{v}|} = \inf_{\boldsymbol{v} \neq \boldsymbol{0}} \frac{|U(\boldsymbol{b})\boldsymbol{v}|}{|T\boldsymbol{v}|} > 1 - \varepsilon, \tag{17.3}$$

$$\sup_{\boldsymbol{v} \neq \boldsymbol{0}} \frac{|D\boldsymbol{h}(\boldsymbol{b})\boldsymbol{v}|}{|T\boldsymbol{v}|} = \sup_{\boldsymbol{v} \neq \boldsymbol{0}} \frac{|U(\boldsymbol{b})\boldsymbol{v}|}{|T\boldsymbol{v}|} < 1 + \varepsilon \tag{17.4}$$



Note that it is not clear whether $\boldsymbol{c} \in E(T, \boldsymbol{c}, i)$ because of the above two requirements. What is going on here is that we are looking for $\boldsymbol{b}$ such that $D\boldsymbol{h}(\boldsymbol{b})$ is sufficiently close to one of those T which also are in a piece of B. Thus we start with one of those T and one of those points $\boldsymbol{c}$ and look for all $\boldsymbol{b}$, if any, which do the right things. In one dimension, the T and $D\boldsymbol{h}(\boldsymbol{b})$ would just be slopes. There are countably many of these pieces of B being denoted as $E(T, \boldsymbol{c}, i)$.

The union of these $E(T, \boldsymbol{c}, i)$ is all of B because if $\boldsymbol{b} \in B$,

$$|\boldsymbol{h}(\boldsymbol{a}) - \boldsymbol{h}(\boldsymbol{b}) - D\boldsymbol{h}(\boldsymbol{b})(\boldsymbol{a} - \boldsymbol{b})| < \varepsilon |U(\boldsymbol{b})(\boldsymbol{a} - \boldsymbol{b})| \tag{17.5}$$

whenever $\boldsymbol{a} \in B\left(\boldsymbol{b}, \frac{2}{i}\right)$ provided i is sufficiently large. Thus also, by Lemma 17.2.1, there is $T \in \mathscr{S}$ such that the above holds for $U(\boldsymbol{b})$ replaced with T and $\boldsymbol{a} \in B\left(\boldsymbol{b}, \frac{2}{i}\right)$ and also 17.3, 17.4. Thus $\boldsymbol{b} \in E(T, \boldsymbol{c}, i)$, so indeed the union of these sets is B.

Now let $\boldsymbol{a}, \boldsymbol{b} \in E(T, \boldsymbol{c}, i)$. Since $\boldsymbol{a}, \boldsymbol{b} \in E(T, \boldsymbol{c}, i)$, $\boldsymbol{a}, \boldsymbol{b}$ are within $1/i$ of $\boldsymbol{c}$ and so $\boldsymbol{a}$ is within $2/i$ of $\boldsymbol{b}$ and so 17.2 holds because of the definition of $E(T, \boldsymbol{c}, i)$. Therefore, from 17.2 and the inequalities which follow, 17.3 and 17.4,

$$(1 - 3\varepsilon)|T(\boldsymbol{a} - \boldsymbol{b})| \leq |\boldsymbol{h}(\boldsymbol{a}) - \boldsymbol{h}(\boldsymbol{b})| \leq (1 + 3\varepsilon)|T(\boldsymbol{a} - \boldsymbol{b})| \tag{17.6}$$

Indeed, from 17.5,

$$\begin{aligned} |\boldsymbol{h}(\boldsymbol{a})-\boldsymbol{h}(\boldsymbol{b})| &< |U(\boldsymbol{b})(\boldsymbol{a}-\boldsymbol{b})|+\varepsilon|U(\boldsymbol{b})(\boldsymbol{a}-\boldsymbol{b})| = (1+\varepsilon)|U(\boldsymbol{b})(\boldsymbol{a}-\boldsymbol{b})| \\ &< (1+\varepsilon)^2|T(\boldsymbol{a}-\boldsymbol{b})| < (1+3\varepsilon)|T(\boldsymbol{a}-\boldsymbol{b})| \end{aligned}$$

The other side of 17.6 is similar.

Note that this proves that on $E(T,\boldsymbol{c},i)$ the function $\boldsymbol{h}$ is one to one and $U(\boldsymbol{b})$ is a close approximation to T for each $\boldsymbol{b}\in E(T,\boldsymbol{c},i)$. It also shows the Lipschitz continuity of $\boldsymbol{h}$ on $E(T,\boldsymbol{c},i)$ and $\boldsymbol{h}^{-1}$ on $\boldsymbol{h}(E(T,\boldsymbol{c},i))$ by comparison with T. What has just been shown is a very interesting result for its own sake. Its implications are in the following lemma.

From 17.6, and letting $\boldsymbol{x},\boldsymbol{y}\in T(E(T,\boldsymbol{c},i))$, It follows that

$$\left|\boldsymbol{h}\left(T^{-1}(\boldsymbol{x})\right)-\boldsymbol{h}\left(T^{-1}(\boldsymbol{y})\right)\right| \le (1+3\varepsilon)|\boldsymbol{x}-\boldsymbol{y}| \tag{17.7}$$

Also if $\boldsymbol{x},\boldsymbol{y}$ in $\boldsymbol{h}(E(T,\boldsymbol{c},i))$, the left inequality in 17.6 implies

$$\left|T\left(\boldsymbol{h}^{-1}(\boldsymbol{x})\right)-T\left(\boldsymbol{h}^{-1}(\boldsymbol{y})\right)\right| \le \frac{1}{(1-3\varepsilon)}|\boldsymbol{x}-\boldsymbol{y}| \tag{17.8}$$

Thus, $\boldsymbol{h}$ is one to one on $E(T,\boldsymbol{c},i)$, $\boldsymbol{h}^{-1}$ is one to one on $\boldsymbol{h}(E(T,\boldsymbol{c},i))$, and both functions are Lipschitz continuous, $\boldsymbol{h}^{-1}$ being Lipschitz continuous because there is some $\delta>0$ such that $|T\boldsymbol{v}|\ge\delta|\boldsymbol{v}|$ due to equivalence of norms and that $\|\boldsymbol{v}\|\equiv|T\boldsymbol{v}|$ is a norm.

17.4 Estimates and a Limit

In what follows, we have G is an open set. For the sake of simplicity, it will always be assumed that continuous $\boldsymbol{h}: G\subseteq\mathbb{R}^n\to\mathbb{R}^m, m\ge n$. There are two parts to the following lemma. In one part, $\boldsymbol{h}$ is not even required to be one to one and in the other part, it is one to one. As above, A is a Borel subset of G such that $\boldsymbol{h}$ is differentiable on A.

Lemma 17.4.1 *Let $\boldsymbol{h}$ be differentiable on $A\subseteq G$, $\boldsymbol{h}$ continuous on G, and let A^+ consist of those points $\boldsymbol{x}\in A$ where $\det\left(D\boldsymbol{h}(\boldsymbol{x})^*D\boldsymbol{h}(\boldsymbol{x})\right)>0$. Then if B is any Borel subset of A^+, there is a disjoint sequence of Borel sets and invertible symmetric transformations T_k, $\{(E_k,T_k)\}$, $\cup_k E_k=B$ such that $\boldsymbol{h}$ is Lipschitz on E_k and $\boldsymbol{h}^{-1}$ is Lipschitz on $\boldsymbol{h}(E_k)$; in fact $\boldsymbol{h}^{-1}$ is Lipschitz on a larger set $\boldsymbol{h}(E(T,\boldsymbol{c},i))$ for suitable $T,\boldsymbol{c},i$. Also for any $\boldsymbol{b}\in E_k$, 17.3 and 17.4 both hold. Thus, for $\boldsymbol{b}\in E_k$*

$$(1-\varepsilon)|T_k\boldsymbol{v}| < |D\boldsymbol{h}(\boldsymbol{b})\boldsymbol{v}| = |U(\boldsymbol{b})\boldsymbol{v}| < (1+\varepsilon)|T_k\boldsymbol{v}| \tag{17.9}$$

One can also conclude that for $\boldsymbol{b}\in E_k$,

$$(1-\varepsilon)^{-n}|\det(T_k)| \le \det(U(\boldsymbol{b})) \le (1+\varepsilon)^n|\det(T_k)| \tag{17.10}$$

Also if $\boldsymbol{h}$ is one to one on $B\subseteq A^+$,

$$\mathscr{H}^n(\boldsymbol{h}(B)) = \int_B \det(U(\boldsymbol{x}))\,dx \tag{17.11}$$

Letting $Z=A\setminus A^+$, suppose $\boldsymbol{h}$ is Lipschitz on G. Then

$$\mathscr{H}^n(\boldsymbol{h}(Z)) = 0. \tag{17.12}$$

Proof: From 17.7,

$$|\boldsymbol{h}(\boldsymbol{a})-\boldsymbol{h}(\boldsymbol{b})| \leq (1+3\varepsilon)\,|T(\boldsymbol{a})-T(\boldsymbol{b})| \leq (1+3\varepsilon)\,\|T\|\,|\boldsymbol{a}-\boldsymbol{b}| \tag{17.13}$$

Using the fact that T is one to one, there is $\delta>0$ such that $|T\boldsymbol{z}| \geq \delta\,|\boldsymbol{z}|$ so 17.8 implies that

$$\left|\boldsymbol{h}^{-1}(\boldsymbol{x})-\boldsymbol{h}^{-1}(\boldsymbol{y})\right| \leq \frac{1}{\delta(1-3\varepsilon)}\,|\boldsymbol{x}-\boldsymbol{y}| \tag{17.14}$$

Now let (E_k, T_k) result from a disjoint union of measurable subsets of the countably many $E(T,\boldsymbol{c},i)$ such that $B=\cup_k E_k$. Thus the above Lipschitz conditions 17.7 and 17.8 hold for T_k in place of T. It is not necessary to assume $\boldsymbol{h}$ is one to one in this part of the lemma. $\boldsymbol{h}^{-1}$ refers to the inverse image of $\boldsymbol{h}$ restricted to $\boldsymbol{h}(E_k)$ as discussed above.

Finally, consider 17.10. 17.9 implies that

$$(1-\varepsilon)\,|\boldsymbol{v}| < \left|U(\boldsymbol{b})\,T_k^{-1}\boldsymbol{v}\right| < (1+\varepsilon)\,|\boldsymbol{v}|$$

A generic vector in $B(\mathbf{0},1-\varepsilon)$ is $(1-\varepsilon)\,\boldsymbol{v}$ where $|\boldsymbol{v}|<1$. Thus, the above inequality implies

$$B(\mathbf{0},1-\varepsilon) \subseteq U(\boldsymbol{b})\,T_k^{-1}B(\mathbf{0},1) \subseteq B(\mathbf{0},1+\varepsilon)$$

This implies

$$\alpha(n)\,(1-\varepsilon)^n \leq \det\left(U(\boldsymbol{b})\,T_k^{-1}\right)\alpha(n) \leq \alpha(n)\,(1+\varepsilon)^n$$

and so

$$(1-\varepsilon)^n \leq \det(U(\boldsymbol{b}))\det\left(T_k^{-1}\right) \leq (1+\varepsilon)^n$$

and so for $\boldsymbol{b}\in E_k$,

$$(1-\varepsilon)^n\,|\det(T_k)| \leq \det(U(\boldsymbol{b})) \leq (1+\varepsilon)^n\,|\det(T_k)| \tag{17.15}$$

It remains to consider 17.11. Recall that B was a Borel measurable subset of A^+ the set where $U(\boldsymbol{x})^{-1}$ exists.

From 17.6, $(1-3\varepsilon)\,|T_k(\boldsymbol{a}-\boldsymbol{b})| \leq |\boldsymbol{h}(\boldsymbol{a})-\boldsymbol{h}(\boldsymbol{b})| \leq (1+3\varepsilon)\,|T_k(\boldsymbol{a}-\boldsymbol{b})|$ and the definition of Hausdorff measure,

$$\begin{aligned}(1-3\varepsilon)\,|\det(T_k)|\,m_n(E_k) &= (1-3\varepsilon)\,\mathscr{H}^n(T_k(E_k)) \leq \mathscr{H}^n(\boldsymbol{h}(E_k)) \\ &\leq (1+3\varepsilon)\,\mathscr{H}^n(T_k(E_k)) = (1+3\varepsilon)\,|\det(T_k)|\,m_n(E_k)\end{aligned}$$

Then

$$(1-3\varepsilon)\int_{E_k}|\det(T_k)|\,dm_n \leq \mathscr{H}^n(\boldsymbol{h}(E_k)) \leq (1+3\varepsilon)\int_{E_k}|\det(T_k)|\,dm_n$$

From 17.15

$$\frac{1-3\varepsilon}{(1+\varepsilon)^n}\int_{E_k}|\det(U(\boldsymbol{b}))|\,dm_n \leq \mathscr{H}^n(\boldsymbol{h}(E_k)) \leq \frac{(1+3\varepsilon)}{(1-\varepsilon)^n}\int_{E_k}|\det(U(\boldsymbol{b}))|\,dm_n$$

Then if we assume $\boldsymbol{h}$ is one to one on B, adding these over all k yields

$$\frac{1-2\varepsilon}{(1-\varepsilon)^n}\int_B|\det U|\,dm_n \leq \mathscr{H}^n(\boldsymbol{h}(B)) \leq \frac{1+2\varepsilon}{(1-\varepsilon)^n}\int_B|\det U|\,dm_n$$

Since ε is arbitrary,

$$\mathscr{H}^n(\boldsymbol{h}(B)) = \int_B |\det U|\, dm_n \tag{17.16}$$

Now consider 17.12 where we assume $\boldsymbol{h}$ is Lipschitz on G. First suppose G is bounded. Let $\boldsymbol{k}_\varepsilon(\boldsymbol{x}) \equiv \begin{pmatrix} \boldsymbol{h}(\boldsymbol{x}) \\ \varepsilon \boldsymbol{x} \end{pmatrix}$ so $\boldsymbol{k}_\varepsilon$ is one to one. Then for all $\boldsymbol{x} \in A$,

$$\det\left(D\boldsymbol{k}_\varepsilon(\boldsymbol{x})^* D\boldsymbol{k}_\varepsilon(\boldsymbol{x})\right) = \det\left(D\boldsymbol{h}(\boldsymbol{x})^* D\boldsymbol{h}(\boldsymbol{x}) + \varepsilon^2 I_n\right) > 0$$

Note that $P\boldsymbol{k}_\varepsilon = \boldsymbol{h}$ and for $\boldsymbol{k}_\varepsilon$, $A = A^+$. Letting $\{E_k\}$ be the disjoint Borel sets on which $\boldsymbol{k}_\varepsilon$ is Lipschitz and one to one with inverse also Lipschitz, it follows from 17.16

$$\mathscr{H}^n(\boldsymbol{k}_\varepsilon(Z \cap E_k)) = \int_{Z\cap E_k} \det\left(D\boldsymbol{h}(\boldsymbol{x})^* D\boldsymbol{h}(\boldsymbol{x}) + \varepsilon^2 I_n\right)^{1/2} dx$$

Also, since $\boldsymbol{h} = P\boldsymbol{k}_\varepsilon$, where P is Lipschitz with Lipschitz constant no more than 1, it follows from Lemma 17.1.2 that

$$\mathscr{H}^n(\boldsymbol{h}(Z \cap E_k)) \leq \int_{Z\cap E_k} \det\left(D\boldsymbol{h}(\boldsymbol{x})^* D\boldsymbol{h}(\boldsymbol{x}) + \varepsilon^2 I_n\right)^{1/2} dx$$

Then, $\boldsymbol{h}(Z) \subseteq \cup_k \boldsymbol{h}(Z \cap E_k)$. Hence,

$$\begin{aligned} \mathscr{H}^n(\boldsymbol{h}(Z)) &\leq \sum_k \mathscr{H}^n(\boldsymbol{h}(Z\cap E_k)) \leq \sum_k \int_{Z\cap E_k} \det\left(D\boldsymbol{h}(\boldsymbol{x})^* D\boldsymbol{h}(\boldsymbol{x}) + \varepsilon^2 I_n\right)^{1/2} dx \\ &= \int_Z \det\left(D\boldsymbol{h}(\boldsymbol{x})^* D\boldsymbol{h}(\boldsymbol{x}) + \varepsilon^2 I_n\right)^{1/2} dx \end{aligned}$$

Since $\boldsymbol{h}$ is assumed Lipschitz, the expression in the integrand is bounded independent of ε and so, since G is bounded, the dominated convergence theorem applies and it follows that

$$\mathscr{H}^n(\boldsymbol{h}(Z)) \leq \int_Z \det\left(D\boldsymbol{h}(\boldsymbol{x})^* D\boldsymbol{h}(\boldsymbol{x})\right)^{1/2} dx = 0 \tag{17.17}$$

In case G is not bounded, apply the above to $G_n \equiv G \cap B(\boldsymbol{0}, n)$ for $n \in \mathbb{N}$. Then pass to a limit. ■

Note how 17.17 gives a version of Sard's theorem. Also it sufficed to be able to use the dominated convergence theorem in the above argument leading to 17.17. Letting $\boldsymbol{h}$ be Lipschitz on G was a convenient way to make this happen. However, you could likely generalize.

17.5 Area Formulas

Lemma 17.5.1 *Let the $\{E_k\}$ be the disjoint Borel sets whose union is A^+ such that $\boldsymbol{h}$ is Lipschitz on E_k and $\boldsymbol{h}^{-1}$ is Lipschitz on $\boldsymbol{h}(E_k)$. Then if $B \subseteq \mathbb{R}^n$ is Lebesgue measurable, it follows that*

$$\mathscr{H}^n(\boldsymbol{h}(B\cap E_k)) = \int_{B\cap E_k} \det(U(\boldsymbol{x}))\, dx \tag{17.18}$$

and $\boldsymbol{h}(B\cap E_k)$ is $\mathscr{H}^n$ measurable. Also, for all $E \subseteq \mathbb{R}^m$ that are $\mathscr{H}^n$ measurable, all integrals make sense and

$$\int_{\boldsymbol{h}(E_k)} \mathscr{X}_E(\boldsymbol{y})\, d\mathscr{H}^n = \int_{E_k} \mathscr{X}_E(\boldsymbol{h}(\boldsymbol{x})) \det(U(\boldsymbol{x}))\, dx \tag{17.19}$$

If $g \geq 0$ and $\boldsymbol{x} \to \boldsymbol{g}(\boldsymbol{x})$ is $\mathscr{H}^n$ measurable, then everything makes sense and

$$\int_{\boldsymbol{h}(E_k)} g(\boldsymbol{y})\, d\mathscr{H}^n = \int_{E_k} g(\boldsymbol{h}(\boldsymbol{x})) \det(U(\boldsymbol{x}))\, dx \tag{17.20}$$

Proof: By 17.16 the conclusion 17.18 is true if B is a Borel set. Let B be Lebesgue measurable. By regularity, there exist Borel sets F, H such that $F \subseteq B \subseteq H$ and $m_n(H \setminus F) = 0$. Then it follows from the fact that $\mathscr{H}^n$ maps sets of measure zero to sets of measure 0 that $\mathscr{H}^n(\boldsymbol{h}(B \cap E_k)) = \mathscr{H}^n(\boldsymbol{h}(F \cap E_k)) = \mathscr{H}^n(\boldsymbol{h}(H \cap E_k))$ and also $\boldsymbol{h}(B \cap E_k)$ is $\mathscr{H}^n$ measurable. Thus

$$\mathscr{H}^n(\boldsymbol{h}(B \cap E_k)) = \mathscr{H}^n(\boldsymbol{h}(F \cap E_k)) = \int_{F \cap E_k} \det(U(\boldsymbol{x}))\, dx = \int_{B \cap E_k} \det(U(\boldsymbol{x}))\, dx$$

Now consider 17.19. First is a claim.

Claim: $\boldsymbol{h}\left(E_k \cap \boldsymbol{h}^{-1}(F)\right) = \boldsymbol{h}(E_k) \cap F$.

Proof of claim: If $\boldsymbol{y} \in \boldsymbol{h}\left(E_k \cap \boldsymbol{h}^{-1}(F)\right)$ then there is $\boldsymbol{x} \in E_k \cap \boldsymbol{h}^{-1}(F)$ such that $\boldsymbol{y} = \boldsymbol{h}(\boldsymbol{x}) \in \boldsymbol{h}(E_k) \cap F$ and so $\boldsymbol{h}\left(E_k \cap \boldsymbol{h}^{-1}(F)\right) \subseteq \boldsymbol{h}(E_k) \cap F$. If $\boldsymbol{y} \in \boldsymbol{h}(E_k) \cap F$ then $\boldsymbol{y} = \boldsymbol{h}(\boldsymbol{x})$ for $\boldsymbol{x} \in E_k$ and also $\boldsymbol{h}(\boldsymbol{x}) \in F$ so $\boldsymbol{x} \in E_k \cap \boldsymbol{h}^{-1}(F)$ and so $\boldsymbol{y} = \boldsymbol{h}(\boldsymbol{x}) \in \boldsymbol{h}\left(E_k \cap \boldsymbol{h}^{-1}(F)\right)$ which shows $\boldsymbol{h}(E_k) \cap F \subseteq \boldsymbol{h}\left(E_k \cap \boldsymbol{h}^{-1}(F)\right)$ and so these sets are equal as claimed.

First assume B is a Borel set. That way $\boldsymbol{h}^{-1}(B)$ will also be a Borel set. Then, by the claim and the first part,

$$\begin{aligned}\int_{\boldsymbol{h}(E_k)} \mathscr{X}_B(\boldsymbol{y})\, d\mathscr{H}^n &= \mathscr{H}^n(B \cap \boldsymbol{h}(E_k)) = \mathscr{H}^n\left(\boldsymbol{h}\left(E_k \cap \boldsymbol{h}^{-1}(B)\right)\right) \\ &= \int_{E_k} \mathscr{X}_{\boldsymbol{h}^{-1}(B)}(\boldsymbol{x}) \det(U(\boldsymbol{x}))\, dx = \int_{E_k} \mathscr{X}_B(\boldsymbol{h}(\boldsymbol{x})) \det(U(\boldsymbol{x}))\, dx\end{aligned}$$

Now consider $\lambda(E) \equiv \mathscr{H}^n(E \cap \boldsymbol{h}(E_k))$ for E an $\mathscr{H}^n$ measurable function. This makes sense and is a measure because $\boldsymbol{h}$ is one to one, $\boldsymbol{h}(E_k)$ is $\mathscr{H}^n$ measurable because $\boldsymbol{h}$ is Lipschitz on E_k and E_k is a Borel set, hence by Lemma 17.1.4 $\boldsymbol{h}(E_k)$ is Lebesgue measurable. It is a finite measure because these E_k are all bounded and

$$\lambda(\mathbb{R}^m) = \mathscr{H}^n(\boldsymbol{h}(E_k)) = \int_{E_k} \det(U(\boldsymbol{x}))\, dx < \infty$$

By Lemma 9.4.4, λ is regular on Borel sets. However, by Theorem 16.7.1, whenever E is a $\mathscr{H}^n$ measurable set, there exists a Borel set F such that $\lambda(E) = \lambda(F)$ and $F \supseteq E$. Therefore, by Lemma 9.4.4, λ is a regular measure and if E is $\mathscr{H}^n$ measurable, there exist F, H with these being Borel sets and such that $F \subseteq E \subseteq H$ and $\lambda(H \setminus F) = 0$. Therefore,

$$\mathscr{X}_F(\boldsymbol{h}(\boldsymbol{x})) \det(U(\boldsymbol{x})) \leq \mathscr{X}_E(\boldsymbol{h}(\boldsymbol{x})) \det(U(\boldsymbol{x})) \leq \mathscr{X}_H(\boldsymbol{h}(\boldsymbol{x})) \det(U(\boldsymbol{x}))$$

and

$$\int_{E_k} \left(\mathscr{X}_H(\boldsymbol{h}(\boldsymbol{x})) \det(U(\boldsymbol{x})) - \mathscr{X}_F(\boldsymbol{h}(\boldsymbol{x})) \det(U(\boldsymbol{x}))\right) dx = 0$$

so, $\mathscr{X}_H(\boldsymbol{h}(\boldsymbol{x})) \det(U(\boldsymbol{x})) - \mathscr{X}_F(\boldsymbol{h}(\boldsymbol{x})) \det(U(\boldsymbol{x})) = 0$ a.e. By completeness of Lebesgue measure, $\boldsymbol{x} \to \mathscr{X}_E(\boldsymbol{h}(\boldsymbol{x})) \det(U(\boldsymbol{x}))\, \mathscr{X}_{E_k}(\boldsymbol{x})$ is Lebesgue measurable and

$$\begin{aligned}\int_{\boldsymbol{h}(E_k)} \mathscr{X}_E(\boldsymbol{y})\, d\mathscr{H}^n &= \int_{\boldsymbol{h}(E_k)} \mathscr{X}_F(\boldsymbol{y})\, d\mathscr{H}^n = \int_{E_k} \mathscr{X}_F(\boldsymbol{h}(\boldsymbol{x})) \det(U(\boldsymbol{x}))\, dx \\ &= \int_{E_k} \mathscr{X}_E(\boldsymbol{h}(\boldsymbol{x})) \det(U(\boldsymbol{x}))\, dx \blacksquare\end{aligned}$$

Now 17.20 follows right away from Theorem 9.1.6, expressing g as the limit of an increasing sequence of $\mathscr{H}^n$ measurable simple functions and using the monotone convergence theorem. ■

Corollary 17.5.2 *If $A \subseteq G \subseteq \mathbb{R}^n$ an open set and if continuous $\boldsymbol{h} : G \to \mathbb{R}^m, m \geq n$, and if A is a Borel set such that $D\boldsymbol{h}(\boldsymbol{x})$ exists for $\boldsymbol{x} \in A$ and if A^+ is the set where $\det\left(D\boldsymbol{h}(\boldsymbol{x})^* D\boldsymbol{h}(\boldsymbol{x})\right) > 0$. Then if $\boldsymbol{h}$ is one to one on A^+, then if $g \geq 0$ and $\mathscr{H}^n$ measurable,*

$$\int_{\boldsymbol{h}(A^+)} g(\boldsymbol{y})\, d\mathscr{H}^n = \int_{A^+} g(\boldsymbol{h}(\boldsymbol{x})) \det(U(\boldsymbol{x}))\, dx$$

Proof: Let the E_k be the disjoint Borel sets whose union is A^+. Then 17.20 holds and, since $\boldsymbol{h}$ is one to one, you can add and obtain the above result. ■

Now with this, we can give the area formula.

Definition 17.5.3

To save on notation, I will denote $\det(U(\boldsymbol{x}))$ as $J_*(\boldsymbol{x})$.

Theorem 17.5.4

Let $\boldsymbol{h}$ be Lipschitz on the open set G with $\boldsymbol{h} : G \to \mathbb{R}^m, m \geq n$ and let A be a Borel set on which $D\boldsymbol{h}$ exists such that $m_n(G \setminus A) = 0$. Let A^+ be those points of A where $J_*(\boldsymbol{x}) > 0$. Also let $\boldsymbol{h}$ be one to one on A^+. Then whenever $g \geq 0, g$ being $\mathscr{H}^n$ measurable,

$$\int_{\boldsymbol{h}(A)} g(\boldsymbol{y})\, d\mathscr{H}^n = \int_A g(\boldsymbol{h}(\boldsymbol{x})) \det(U(\boldsymbol{x}))\, dx$$

and the integrals all make sense.

Proof: By Corollary 17.5.2 and 17.12 of Lemma 17.4.1,

$$\begin{aligned} \int_{\boldsymbol{h}(A)} g(\boldsymbol{y})\, d\mathscr{H}^n &= \int_{\boldsymbol{h}(A^+)} g(\boldsymbol{y})\, d\mathscr{H}^n = \int_{A^+} g(\boldsymbol{h}(\boldsymbol{x})) J_*(\boldsymbol{x})\, dx \\ &= \int_A g(\boldsymbol{h}(\boldsymbol{x})) J_*(\boldsymbol{x})\, dx \blacksquare \end{aligned}$$

Note that this is essentially the same formula presented earlier in Corollary 14.3.5 if you identify the surface measure σ_n with $\mathscr{H}^n$ which is implied by the above formula. The difference here is that the surface measure is obviously an intrinsic property of $\boldsymbol{h}(A)$ which has geometric significance whereas in the earlier treatment, I pulled $J_*(\boldsymbol{x})$ out of the air based solely on its resemblance to something in beginning calculus. Of course a more complete geometric justification is easily available for volumes of parallelepiped, but this evades the question whether this information about parallelepipeds would be meaningful in the context of a Lipschitz manifold. It is also simpler here because it was not necessary to consider a partition of unity.

Next we consider the case where $\boldsymbol{h}$ is perhaps not one to one but is Lipschitz on G.

Definition 17.5.5

Letting A be as in Theorem 17.5.4, let $\#(\boldsymbol{y})$ denote the number of $\boldsymbol{x} \in A$ in $\boldsymbol{h}^{-1}(\boldsymbol{y})$.

Letting A^+ be the set of $\boldsymbol{x} \in A$ for which $J_*(\boldsymbol{x}) > 0$ as above, also let $\{E_k\}$ be the disjoint Borel sets whose union is all of A^+ on each of which $\boldsymbol{h}$ is one to one and Lipschitz, let $n(\boldsymbol{y}) \equiv \sum_k \mathscr{X}_{\boldsymbol{h}(E_k)}(\boldsymbol{y})$. Thus $n(\boldsymbol{y}) = \#(\boldsymbol{y})$ on $\boldsymbol{h}(A^+)$ so these two functions coincide everywhere except for a set of $\mathscr{H}^n$ measure zero $\boldsymbol{h}(A \setminus A^+)$. Clearly $n(\boldsymbol{y})$ is measurable

because $\boldsymbol{h}$ is Lipschitz on E_k which is Borel measurable and Lipschitz maps take Lebesgue measurable sets to $\mathscr{H}^n$ measurable sets. Therefore, # is also $\mathscr{H}^n$ measurable because $\mathscr{H}^n$ is complete. Now from Lemma 17.5.1, if $g \geq 0$ and g is $\mathscr{H}^n$ measurable,

$$\int_{\boldsymbol{h}(A^+)} \mathscr{X}_{\boldsymbol{h}(E_k)}(\boldsymbol{y})\, g(\boldsymbol{y})\, d\mathscr{H}^n = \int_{E_k} g(\boldsymbol{h}(\boldsymbol{x}))\, J_*(\boldsymbol{x})\, dx$$

Now add these and use the monotone convergence theorem as required along with the fact that $\boldsymbol{h}(Z)$ has $\mathscr{H}^n$ measure 0.

$$\begin{aligned}\int_{\boldsymbol{h}(A)} \#(\boldsymbol{y})\, g(\boldsymbol{y})\, d\mathscr{H}^n &= \int_{\boldsymbol{h}(A^+)} n(\boldsymbol{y})\, g(\boldsymbol{y})\, d\mathscr{H}^n = \int_{\boldsymbol{h}(A^+)} \sum_k \mathscr{X}_{\boldsymbol{h}(E_k)}(\boldsymbol{y})\, g(\boldsymbol{y})\, d\mathscr{H}^n \\ &= \int_{A^+} g(\boldsymbol{h}(\boldsymbol{x}))\, J_*(\boldsymbol{x})\, dx = \int_A g(\boldsymbol{h}(\boldsymbol{x}))\, J_*(\boldsymbol{x})\, dx.\end{aligned}$$

This takes care of the case where $\boldsymbol{h}$ is not one to one. This version is in the following theorem.

Theorem 17.5.6 *Suppose $\boldsymbol{h}: G \to \mathbb{R}^m$ is Lipschitz and let A be a Borel set on which $D\boldsymbol{h}$ exists with $m_p(G \setminus A) = 0$. Then if $g \geq 0$ and is $\mathscr{H}^n$ measurable,*

$$\int_{\boldsymbol{h}(A)} \#(\boldsymbol{y})\, g(\boldsymbol{y})\, d\mathscr{H}^n = \int_A g(\boldsymbol{h}(\boldsymbol{x}))\, J_*(\boldsymbol{x})\, dx.$$

and everything makes sense where here $\#(\boldsymbol{y})$ is defined in Definition 17.5.5 as the number of times $\boldsymbol{h}$ hits $\boldsymbol{y}$ from points in A.

One other observation might be useful. What if you simply defined $\hat{A}$ to be those points $\boldsymbol{x}$ where $D\boldsymbol{h}(\boldsymbol{x})$ exists. It might not be obvious that this set is Borel. However, it is Lebesgue measurable, since by Rademacher's theorem it differs from G by a set of measure zero and so by regularity of the measure, it can be approximated by an F_σ set A such that $m_n(\hat{A} \setminus A) = 0$. Thus A would work in all of the above and one could replace A wherever it occurs in Theorems 17.5.6 and 17.5.4 with $\hat{A}$ and the conclusions of these theorems would be unchanged.

Next is an interesting version of the chain rule for Lipschitz maps.

Theorem 17.5.7 *Let $\boldsymbol{f}$, $\boldsymbol{g}$ be Lipschitz mappings from $\mathbb{R}^n$ to $\mathbb{R}^n$ with $\boldsymbol{f}(\boldsymbol{g}(\boldsymbol{x})) = \boldsymbol{x}$ on A, a measurable set. Then for a.e. $\boldsymbol{x} \in A$, $D\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}))$, $D\boldsymbol{f}(\boldsymbol{x})$, and $D(\boldsymbol{f} \circ \boldsymbol{g})(\boldsymbol{x})$ all exist and $I = D(\boldsymbol{g} \circ \boldsymbol{f})(\boldsymbol{x}) = D\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}))\, D\boldsymbol{f}(\boldsymbol{x})$.*

The proof of this theorem is based on the following lemma.

Lemma 17.5.8 *If $\boldsymbol{h}: \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz, then if $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}$ for all $\boldsymbol{x} \in A$, then*

$$\det(D\boldsymbol{h}(\boldsymbol{x})) = 0 \text{ a.e.} \boldsymbol{x} \in A.$$

Proof: By the Area formula, $0 = \int_{\{\boldsymbol{0}\}} \#(\boldsymbol{y})\, dy = \int_A |\det(D\boldsymbol{h}(\boldsymbol{x}))|\, dx$, and so it follows that $\det(D\boldsymbol{h}(\boldsymbol{x})) = 0$ *a.e.* ■

Proof of the theorem: On A, $\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x})) - \boldsymbol{x} = \boldsymbol{0}$ and so by the lemma, there exists a set of measure zero, N_1 such that if $\boldsymbol{x} \notin N_1$, $D(\boldsymbol{g} \circ \boldsymbol{f})(\boldsymbol{x}) - I = \boldsymbol{0}$. Let M be the set of points in $\boldsymbol{f}(\mathbb{R}^n)$ where $\boldsymbol{g}$ fails to be differentiable and let $N_2 \equiv \boldsymbol{g}(M) \cap A$, also a set of measure zero. Finally let N_3 be the set of points where $\boldsymbol{f}$ fails to be differentiable. Then if $\boldsymbol{x} \notin N_1 \cup N_2 \cup N_3$, the chain rule implies $I = D(\boldsymbol{g} \circ \boldsymbol{f})(\boldsymbol{x}) = D\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}))\, D\boldsymbol{f}(\boldsymbol{x})$. ■

17.6 The Divergence Theorem

Using Rademacher's theorem, all conditions are satisfied for Definition 14.4.1 provided each of the g_i used there are Lipschitz. When this happens, we say U is a bounded open set with a Lipschitz boundary which lies on one side of its boundary. Thus we obtain the divergence theorem. Here I will use $\mathscr{H}^{p-1}$ for surface measure σ on the boundary of U since, by the area formula, this is what it is.

Theorem 17.6.1 *Let U be a bounded open set with a Lipschitz boundary which lies on one side of its boundary. Then if $f \in C_c^1(\mathbb{R}^p)$,*

$$\int_U f_{,k}(\boldsymbol{x})\, dm_p = \int_{\partial U} f n_k d\mathscr{H}^{p-1} \tag{17.21}$$

where $\boldsymbol{n} = (n_1, \cdots, n_n)$ is the $\mathscr{H}^{p-1}$ measurable unit outer normal. Also, if $\boldsymbol{F}$ is a vector field such that each component is in $C_c^1(\mathbb{R}^p)$, then

$$\int_U \nabla \cdot \boldsymbol{F}(\boldsymbol{x})\, dm_p = \int_{\partial U} \boldsymbol{F} \cdot \boldsymbol{n} d\mathscr{H}^{p-1}. \tag{17.22}$$

Proof: To obtain the first formula from the second which was proved earlier, consider

$$\boldsymbol{F} \equiv \begin{pmatrix} 0 & \cdots & f & \cdots & 0 \end{pmatrix}^T$$

where f is in the k^{th} slot. ■

You could approximate $\boldsymbol{F}$ in the above theorem by convolving with a mollifier as in Lemma 16.4.1 yielding a modified $\boldsymbol{F}$ with one in which each component is in infinitely differentiable, apply the divergence theorem of Theorem 14.4.3 to these and pass to a limit using the dominated convergence theorem to obtain the divergence theorem for $\boldsymbol{F}$. Thus the following corollary is obtained.

Corollary 17.6.2 *In the context of Theorem 17.6.1 it suffices to assume $\boldsymbol{F}$ is Lipschitz.*

17.7 The Reynolds Transport Formula

The Reynolds transport formula is a generalization of the formula for taking the derivative under an integral. It depends on the divergence theorem. I will use the chain rule of Theorem 17.5.7 as needed.

$$\frac{d}{dt}\int_{a(t)}^{b(t)} f(x,t)\, dx = \int_{a(t)}^{b(t)} \frac{\partial f}{\partial t}(x,t)\, dx + f(b(t),t)\, b'(t) - f(a(t),t)\, a'(t)$$

First is an interesting lemma about the determinant. A $p \times p$ matrix can be thought of as a vector in $\mathbb{C}^{p^2}$. Just imagine stringing it out into one long list of numbers. In fact, a way to give the norm of a matrix is just $\sum_i \sum_j |A_{ij}|^2 \equiv \|A\|^2$. This is called the Frobenius norm for a matrix. It makes no difference since all norms are equivalent, but this one is convenient in what follows. Also recall that det maps $p \times p$ matrices to $\mathbb{C}$. It makes sense to ask for the derivative of det on the set of invertible matrices, an open subset of $\mathbb{C}^{p^2}$ with the norm measured as just described because $A \to \det(A)$ is continuous, so the set where $\det(A) \neq 0$ would be an open set. Recall from linear algebra that the sum of the entries

on the main diagonal satisfies $\text{trace}(AB) = \text{trace}(BA)$ whenever both products make sense. Indeed, $\text{trace}(AB) \equiv \sum_i \sum_j A_{ij} B_{ji} = \text{trace}(BA)$

This next lemma is a very interesting observation about the determinant of a matrix added to the identity.

Lemma 17.7.1 $\det(I+U) = 1 + \text{trace}(U) + o(U)$ *where* $o(U)$ *is defined in terms of the Frobenius norm for* $p \times p$ *matrices.*

Proof:

$$\det(I+U) \equiv \sum_{i_1, i_2, \ldots, i_p} \text{sgn}(i_1, i_2, \ldots, i_p) (\delta_{i_1 1} + U_{i_1 1}) \cdots (\delta_{i_p p} + U_{i_p p})$$

which equals $\det(I)$ added to $\text{trace}(U)$ added to a sum of higher order terms of products of the U_{ij}. The $\text{trace}(U)$ comes from using only one U_{ij} in the above product. The resulting term will be 0 unless $i = j$ and so the end result of these will be $\text{trace}(U)$. Of course if more of the U_{ij} are included in the product, this yields the $o(U)$ term. ■

Of course, by equivalence of norms, one could use any other norm for the $p \times p$ matrices.

With this lemma, it is easy to find $D\det(F)$ whenever F is invertible.

$$\begin{aligned} \det(F+U) &= \det\left(F\left(I + F^{-1}U\right)\right) = \det(F)\det\left(I + F^{-1}U\right) \\ &= \det(F)\left(1 + \text{trace}\left(F^{-1}U\right) + o(U)\right) \\ &= \det(F) + \det(F)\,\text{trace}\left(F^{-1}U\right) + o(U) \end{aligned}$$

Therefore, $\det(F+U) - \det(F) = \det(F)\,\text{trace}\left(F^{-1}U\right) + o(U)$. This proves the following.

Proposition 17.7.2 *Let* F^{-1} *exist. Then* $D\det(F)(U) = \det(F)\,\text{trace}\left(F^{-1}U\right)$.

From this, suppose $F(t)$ is a $p \times p$ matrix and all entries are differentiable. Then the following describes $\frac{d}{dt}\det(F)(t)$.

Proposition 17.7.3 *Let* $F(t)$ *be a* $p \times p$ *matrix and all entries are at least Lipschitz. Then for a.e.* t

$$\frac{d}{dt}\det(F)(t) = \det(F(t))\,\text{trace}\left(F^{-1}(t)F'(t)\right) = \det(F(t))\,\text{trace}\left(F'(t)F^{-1}(t)\right) \quad (17.23)$$

Proof: From the above,

$$\begin{aligned} & \det(F(t+h)) - \det(F(t)) \\ = \; & \det(F(t))\,\text{trace}\left(F^{-1}(F(t+h) - F(t))\right) + \overbrace{o(F(t+h) - F(t))}^{=o(h) \text{ since } F' \text{ exists}} \end{aligned}$$

Dividing by h and taking a limit yields 17.23. ■

Let $\boldsymbol{y} = \boldsymbol{h}(t, \boldsymbol{x})$ with $F = F(t, \boldsymbol{x}) = D_2 \boldsymbol{h}(t, \boldsymbol{x})$. I will write $\nabla_{\boldsymbol{y}}$ to indicate the gradient with respect to the $\boldsymbol{y}$ variables and F' to indicate $\frac{\partial}{\partial t} F(t, \boldsymbol{x})$. I will be assuming what is needed to use the various theorems. In particular let $\boldsymbol{h}$ be differentiable and one to one in $\boldsymbol{x}$. Note that $\boldsymbol{h}(t, \boldsymbol{x}) = \boldsymbol{y}$ and so by the inverse function theorem, or actually Corollary 8.10.6,

this defines $\boldsymbol{x}$ as a function of $\boldsymbol{y}$, also differentiable as $\boldsymbol{h}$ because it is always assumed $\det F > 0$.

Now let V_t be $\boldsymbol{h}(t, V_0)$ where V_0 is an open bounded set. Let V_0 have a Lipschitz boundary so one can use the divergence theorem on V_0. Thus this is concerned with smooth motion of a bounded open set with Lipschitz boundary. Let $(t, \boldsymbol{y}) \to \boldsymbol{f}(t, \boldsymbol{y})$ be Lipschitz. The idea is to simplify $\frac{d}{dt}\int_{V_t} \boldsymbol{f}(t, \boldsymbol{y})\, dm_p(y)$. This will involve the change of variables in which the Jacobian will be $\det(F)$ which is assumed positive thus preserving the orientation of the normal vector for V_0 and V_t. In applications of this theory, $\det(F) \leq 0$ is not physically possible. Since $\boldsymbol{h}(t, \cdot)$ is better than Lipschitz and the boundary of V_0 is Lipschitz, V_t will be such that one can use the divergence theorem because the composition of Lipschitz functions is Lipschitz. Then, using the dominated convergence theorem as needed along with the area formula,

$$\frac{d}{dt}\int_{V_t} \boldsymbol{f}(t, \boldsymbol{y})\, dm_p(y) = \frac{d}{dt}\int_{V_0} \boldsymbol{f}(t, \boldsymbol{h}(t, \boldsymbol{x}))\det(F)\, dm_p(x) \tag{17.24}$$

$$= \int_{V_0} \frac{\partial}{\partial t}\boldsymbol{f}(\cdot, \boldsymbol{h}(\cdot, \boldsymbol{x}))\det(F)\, dm_p(x) + \int_{V_0} \boldsymbol{f}(t, \boldsymbol{h}(t, \boldsymbol{x}))\frac{\partial}{\partial t}(\det(F))\, dm_p(x)$$

$$= \int_{V_0} \frac{\partial}{\partial t}(\boldsymbol{f}(t, \boldsymbol{h}(t, \boldsymbol{x})))\det(F)\, dm_p(x) + \int_{V_0} \boldsymbol{f}(t, \boldsymbol{h}(t, \boldsymbol{x}))\operatorname{trace}\left(F'F^{-1}\right)\det(F)\, dm_p(x)$$

$$\begin{aligned} &= \int_{V_0} \left(\frac{\partial}{\partial t}\boldsymbol{f}(t, \boldsymbol{h}(t, \boldsymbol{x})) + \sum_i \frac{\partial \boldsymbol{f}}{\partial y_i}\frac{\partial y_i}{\partial t}\right)\det(F)\, dm_p(x) \\ &\quad + \int_{V_0} \boldsymbol{f}(t, \boldsymbol{h}(t, \boldsymbol{x}))\operatorname{trace}\left(F'F^{-1}\right)\det(F)\, dm_p(x) \end{aligned}$$

$$= \int_{V_t} \frac{\partial}{\partial t}\boldsymbol{f}(t, \boldsymbol{y})\, dm_p(y) + \int_{V_t} \sum_i \frac{\partial \boldsymbol{f}}{\partial y_i}\frac{\partial y_i}{\partial t} + \boldsymbol{f}(t, \boldsymbol{y})\operatorname{trace}\left(F'F^{-1}\right) dm_p(y)$$

Now $\boldsymbol{v} \equiv \frac{\partial}{\partial t}\boldsymbol{h}(t, \boldsymbol{x})$ and also, as noted above, $\boldsymbol{y} \equiv \boldsymbol{h}(t, \boldsymbol{x})$ defines $\boldsymbol{y}$ as a function of $\boldsymbol{x}$ and so $\operatorname{trace}\left(F'F^{-1}\right) = \sum_\alpha \frac{\partial v_i}{\partial x_\alpha}\frac{\partial x_\alpha}{\partial y_i}$. Hence the double sum $\sum_{\alpha,i} \frac{\partial v_i}{\partial x_\alpha}\frac{\partial x_\alpha}{\partial y_i}$ is $\frac{\partial v_i}{\partial y_i} = \nabla_{\boldsymbol{y}} \cdot \boldsymbol{v}$. The above then gives

$$\int_{V_t} \frac{\partial}{\partial t}\boldsymbol{f}(t, \boldsymbol{y})\, dm_p(y) + \int_{V_t} \left(\sum_i \frac{\partial \boldsymbol{f}}{\partial y_i}\frac{\partial y_i}{\partial t} + \boldsymbol{f}(t, \boldsymbol{y})\nabla_{\boldsymbol{y}} \cdot \boldsymbol{v}\right) dm_p(y)$$

$$= \int_{V_t} \frac{\partial}{\partial t}\boldsymbol{f}(t, \boldsymbol{y})\, dm_p(y) + \int_{V_t} (D_2\boldsymbol{f}(t, \boldsymbol{y})\boldsymbol{v} + \boldsymbol{f}(t, \boldsymbol{y})\nabla_{\boldsymbol{y}} \cdot \boldsymbol{v})\, dm_p(y) \tag{17.25}$$

Now consider the i^{th} component of the second integral in the above. It is

$$\int_{V_t} \nabla_{\boldsymbol{y}} f_i(t, \boldsymbol{y}) \cdot \boldsymbol{v} + f_i(t, \boldsymbol{y})\nabla_{\boldsymbol{y}} \cdot \boldsymbol{v} dm_p(y) = \int_{V_t} \nabla_{\boldsymbol{y}} \cdot (f_i(t, \boldsymbol{y})\boldsymbol{v})\, dm_p(y)$$

At this point, use the divergence theorem to get this equals $\int_{\partial V_t} f_i(t, \boldsymbol{y})\boldsymbol{v} \cdot \boldsymbol{n} d\mathscr{H}^{p-1}$. Therefore, from 17.25 and 17.24,

$$\frac{d}{dt}\int_{V_t} \boldsymbol{f}(t, \boldsymbol{y})\, dm_p(y) = \int_{V_t} \frac{\partial}{\partial t}\boldsymbol{f}(t, \boldsymbol{y})\, dm_p(y) + \int_{\partial V_t} \boldsymbol{f}(t, \boldsymbol{y})\boldsymbol{v} \cdot \boldsymbol{n} d\mathscr{H}^{p-1} \tag{17.26}$$

this is the Reynolds transport formula.

Proposition 17.7.4 *Let $\boldsymbol{y} = \boldsymbol{h}(t,\boldsymbol{x})$ where $\boldsymbol{h}$ is C^1 and let $\boldsymbol{f}$ be Lipschitz continuous and let $V_t \equiv \boldsymbol{h}(t,V_0)$ where V_0 is a bounded open set which is on one side of a Lipschitz boundary so that the divergence theorem holds for V_0. Then 17.26 is obtained.*

As with the divergence theorem, Some generalization should be possible to the case where $\boldsymbol{h}$ giving the motion is only Lipschitz by using the version of the chain rule in Theorem 17.5.7 in the above argument when needed.

17.8 The Coarea Formula

The area formula was discussed above. This formula implies that for E a measurable set

$$\mathscr{H}^n(\boldsymbol{f}(E)) = \int \mathscr{X}_E(\boldsymbol{x}) J_*(\boldsymbol{x})\, dm$$

where $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ for $\boldsymbol{f}$ a Lipschitz mapping and $m \geq n$. It is a version of the change of variables formula for multiple integrals. The coarea formula is a statement about the Hausdorff measure of a set which involves the inverse image of $\boldsymbol{f}$. It is somewhat reminiscent of Fubini's theorem because it writes an integral over $\mathbb{R}^n$ as an iterated integral. However, in this case, the inside integral is with respect to Hausdorff measure taken over a level set rather than some $\mathbb{R}^m$. We will let $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ where $m \leq n$ in what follows.

It is possible to obtain the coarea formula as a computation involving the area formula and some simple linear algebra and this is the approach taken here. I found this formula in [17]. This is a good place to obtain a slightly different proof. This argument follows [36] which came from [17]. I find this material very hard, so I hope what follows doesn't have grievous errors. I have never had occasion to use this coarea formula, but I think it is obviously of enormous significance and gives a very interesting geometric assertion. I will use the form of the chain rule in Theorem 17.5.7 as needed.

To begin with we give the linear algebra identity which will be used. Recall that for a real matrix A^* is just the transpose of A. Thus AA^* and A^*A are symmetric.

Theorem 17.8.1 *Let A be an $m \times n$ matrix and let B be an $n \times m$ matrix for $m \leq n$. Then for I an appropriate size identity matrix,* $\det(I+AB) = \det(I+BA)$.

Proof: Use block multiplication to write

$$\begin{pmatrix} I+AB & 0 \\ B & I \end{pmatrix} \begin{pmatrix} I & A \\ 0 & I \end{pmatrix} = \begin{pmatrix} I+AB & A+ABA \\ B & BA+I \end{pmatrix}$$

$$\begin{pmatrix} I & A \\ 0 & I \end{pmatrix} \begin{pmatrix} I & 0 \\ B & I+BA \end{pmatrix} = \begin{pmatrix} I+AB & A+ABA \\ B & I+BA \end{pmatrix}$$

Hence

$$\begin{pmatrix} I+AB & 0 \\ B & I \end{pmatrix} \begin{pmatrix} I & A \\ 0 & I \end{pmatrix} = \begin{pmatrix} I & A \\ 0 & I \end{pmatrix} \begin{pmatrix} I & 0 \\ B & I+BA \end{pmatrix}$$

so

$$\begin{pmatrix} I & A \\ 0 & I \end{pmatrix}^{-1} \begin{pmatrix} I+AB & 0 \\ B & I \end{pmatrix} \begin{pmatrix} I & A \\ 0 & I \end{pmatrix} = \begin{pmatrix} I & 0 \\ B & I+BA \end{pmatrix}$$

which shows that the two matrices

$$\begin{pmatrix} I+AB & 0 \\ B & I \end{pmatrix}, \begin{pmatrix} I & 0 \\ B & I+BA \end{pmatrix}$$

are similar and so they have the same determinant. Thus $\det(I+AB)=\det(I+BA)$. Note that the two matrices are different sizes. ■

With these lemmas it is now possible to establish the coarea formula. First we define $\Lambda(n,m)$ as all possible ordered lists of m numbers taken from $\{1,2,...,n\}$. Recall $\boldsymbol{x}\in\mathbb{R}^n$ and $\boldsymbol{f}(\boldsymbol{x})\in\mathbb{R}^m$ where $m\leq n$. Recall that this was part of the Binet Cauchy theorem, Theorem 1.9.14,

$$\det\left(D\boldsymbol{f}(\boldsymbol{x})D\boldsymbol{f}(\boldsymbol{x})^*\right)=\sum_{\boldsymbol{i}\in\Lambda(n,m)}\left(\det D_{\boldsymbol{x}_{\boldsymbol{i}}}\boldsymbol{f}(\boldsymbol{x})\right)^2$$

Now let $\boldsymbol{i}_c\in\Lambda(n,n-m)$ consist of the remaining indices taken in order where $\boldsymbol{i}\in\Lambda(n,m)$. For $\boldsymbol{i}=(i_1,\cdots,i_m)$, define $\boldsymbol{x}_{\boldsymbol{i}}\equiv(x_{i_1},...,x_{i_m})$ and $\boldsymbol{x}_{\boldsymbol{i}_c}$ to be the other components of $\boldsymbol{x}$ taken in order. Then let

$$\boldsymbol{f}^{\boldsymbol{i}}(\boldsymbol{x})\equiv\begin{pmatrix}\boldsymbol{f}(\boldsymbol{x})\\ \boldsymbol{x}_{\boldsymbol{i}_c}\end{pmatrix}$$

Thus there are $C(n,n-m)=D(n,m)$ different $\boldsymbol{f}^{\boldsymbol{i}}$.

Example 17.8.2 *Say $\boldsymbol{f}:\mathbb{R}^4\to\mathbb{R}^2$. Here are some examples for $\boldsymbol{f}^{\boldsymbol{i}}$:*

$$\begin{pmatrix}f_1(x_1,x_2,x_3,x_4)\\ f_2(x_1,x_2,x_3,x_4)\\ x_2\\ x_4\end{pmatrix},\begin{pmatrix}f_1(x_1,x_2,x_3,x_4)\\ f_2(x_1,x_2,x_3,x_4)\\ x_1\\ x_2\end{pmatrix},\begin{pmatrix}f_1(x_1,x_2,x_3,x_4)\\ f_2(x_1,x_2,x_3,x_4)\\ x_3\\ x_4\end{pmatrix}$$

Thus $\boldsymbol{f}^{\boldsymbol{i}}:\mathbb{R}^n\to\mathbb{R}^n$. For example, if $\boldsymbol{i}$ consists of the first m of these indices, you have

$$D\boldsymbol{f}^{\boldsymbol{i}}(\boldsymbol{x})=\begin{pmatrix}D_{\boldsymbol{x}_{\boldsymbol{i}}}\boldsymbol{f}(\boldsymbol{x}) & *\\ 0 & I\end{pmatrix}$$

and so

$$\det D\boldsymbol{f}^{\boldsymbol{i}}(\boldsymbol{x})=\det D_{\boldsymbol{x}_{\boldsymbol{i}}}\boldsymbol{f}(\boldsymbol{x}). \tag{17.27}$$

Earlier with the area formula, we integrated $J_*(\boldsymbol{x})\equiv\det\left(D\boldsymbol{f}(\boldsymbol{x})^*D\boldsymbol{f}(\boldsymbol{x})\right)^{1/2}$. With the coarea formula, we integrate $J^*(\boldsymbol{x})\equiv\det\left(D\boldsymbol{f}(\boldsymbol{x})D\boldsymbol{f}(\boldsymbol{x})^*\right)^{1/2}$. This proof involves doing this integration and seeing what happens.

Theorem 17.8.3

Let $\boldsymbol{f}:\mathbb{R}^n\to\mathbb{R}^m$ be a Lipschitz map. Let A be a Lebesgue measurable set in $\mathbb{R}^n$. Then the following formula holds along with all measurability assertions needed for it to make sense.

$$\int_{\mathbb{R}^m}\mathcal{H}^{n-m}\left(A\cap\boldsymbol{f}^{-1}(\boldsymbol{y})\right)d\boldsymbol{y}=\int_A J^*(\boldsymbol{x})\,d\boldsymbol{x} \tag{17.28}$$

where $J^*(\boldsymbol{x})\equiv\det\left(D\boldsymbol{f}(\boldsymbol{x})D\boldsymbol{f}(\boldsymbol{x})^*\right)^{1/2}$.

Proof: First assume A is Borel. Now note that

$$\det\left(D\boldsymbol{f}(\boldsymbol{x})D\boldsymbol{f}(\boldsymbol{x})^*\right)=\sum_{\boldsymbol{i}\in\Lambda(n,m)}\det\left(D\boldsymbol{f}^{\boldsymbol{i}}(\boldsymbol{x})\right)^2$$

by the Binet Cauchy theorem and 17.27. For each $\boldsymbol{i}$, $\boldsymbol{f}^{\boldsymbol{i}}\left(\left\{\boldsymbol{x}:\ \det\left(D\boldsymbol{f}^{\boldsymbol{i}}(\boldsymbol{x})\right)=0\right\}\right)$ has measure zero due to Sard's Lemma, Lemma 11.14.4. For $N\equiv\{\boldsymbol{x}:D\boldsymbol{f}(\boldsymbol{x})\text{ does not exist}\}$,

it follows from Rademacher's theorem that $m_n(N)=0$ so $m_n\left(\boldsymbol{f}^{\boldsymbol{i}}(N)\right)=0$ since $\boldsymbol{f}^{\boldsymbol{i}}$ is Lipschitz. Let A be a Borel subset of $\mathbb{R}^n$ and let

$$A_{\boldsymbol{i}}^{+}\equiv\left\{\boldsymbol{x}\in A:\det\left(D\boldsymbol{f}^{\boldsymbol{i}}(\boldsymbol{x})\right)\neq 0\right\}\setminus N$$

This is the same as the set where $\det\left(D\boldsymbol{f}^{\boldsymbol{i}}(\boldsymbol{x})^{*}D\boldsymbol{f}^{\boldsymbol{i}}(\boldsymbol{x})\right)>0$. Now By Lemma 17.4.1, there exist disjoint Borel measurable sets $\left\{F_j^{\boldsymbol{i}}\right\}_{j=1}^{\infty}$ such that $\boldsymbol{f}^{\boldsymbol{i}}$ is one to one on $F_j^{\boldsymbol{i}}$, $\left(\boldsymbol{f}^{\boldsymbol{i}}\right)^{-1}$ is Lipschitz on $\boldsymbol{f}^{\boldsymbol{i}}\left(F_j^{\boldsymbol{i}}\right)$, and

$$\cup_{j=1}^{\infty}F_j^{\boldsymbol{i}}=\left\{\boldsymbol{x}:D\boldsymbol{f}^{\boldsymbol{i}}(\boldsymbol{x})\text{ exists and}\det\left(D\boldsymbol{f}^{\boldsymbol{i}}(\boldsymbol{x})\right)\neq 0\right\}.$$

If $\boldsymbol{x}\in A_{\boldsymbol{i}}^{+}$, it follows $\boldsymbol{x}\in F_j^{\boldsymbol{i}}$ for some $\boldsymbol{i}$ and j. Hence $\cup_{\boldsymbol{i},j}F_j^{\boldsymbol{i}}=A_{\boldsymbol{i}}^{+}$. Thus $\boldsymbol{f}^{\boldsymbol{i}}\left(A_{\boldsymbol{i}}\setminus A_{\boldsymbol{i}}^{+}\right)$ has measure zero.

Now let $\left\{E_j^{\boldsymbol{i}}\right\}$ be measurable sets such that $E_j^{\boldsymbol{i}}\subseteq F_j^{\boldsymbol{i}}$ for some j, the sets $E_j^{\boldsymbol{i}}$ are disjoint, and their union coincides with $\cup_{\boldsymbol{i},j}F_j^{\boldsymbol{i}}$. Let $\boldsymbol{g}:\mathbb{R}^n\to\mathbb{R}^n$ be a Lipschitz function which equals $\left(\boldsymbol{f}^{\boldsymbol{i}}\right)^{-1}$ on $\boldsymbol{f}^{\boldsymbol{i}}\left(E_j^{\boldsymbol{i}}\right)$. I am supressing the dependence on $\boldsymbol{i}$. Then for any $\boldsymbol{x}\in E_j^{\boldsymbol{i}}$, $\boldsymbol{g}\left(\boldsymbol{f}^{\boldsymbol{i}}(\boldsymbol{x})\right)=\boldsymbol{x}$. In particular, $\boldsymbol{g}_{\boldsymbol{i}_c}\left(\boldsymbol{f}^{\boldsymbol{i}}(\boldsymbol{x})\right)=\boldsymbol{x}_{\boldsymbol{i}_c}$ where

$$\boldsymbol{g}_{\boldsymbol{i}}(\boldsymbol{y})\equiv\left(\begin{array}{ccc} g_{i_1}(\boldsymbol{y}) & \cdots & g_{i_m}(\boldsymbol{y})\end{array}\right)^{T}$$

for $\boldsymbol{i}\equiv(i_1,\cdots,i_m)$ with $\boldsymbol{g}_{\boldsymbol{i}_c}(\boldsymbol{y})$ defined similarly and $\boldsymbol{x}\in E_j^{\boldsymbol{i}}$, with

$$\boldsymbol{y}\equiv\left(\begin{array}{c}\boldsymbol{y}_1\\ \boldsymbol{y}_2\end{array}\right)\equiv\left(\begin{array}{c}\boldsymbol{f}(\boldsymbol{x})\\ \boldsymbol{x}_{\boldsymbol{i}_c}\end{array}\right)\equiv\boldsymbol{f}^{\boldsymbol{i}}(\boldsymbol{x})\in\boldsymbol{f}^{\boldsymbol{i}}\left(E_j^{\boldsymbol{i}}\right),$$

$$\boldsymbol{x}_{\boldsymbol{i}}=\boldsymbol{g}_{\boldsymbol{i}}\left(\boldsymbol{f}^{\boldsymbol{i}}(\boldsymbol{x})\right),\ \boldsymbol{y}_2\equiv\boldsymbol{x}_{\boldsymbol{i}_c}=\boldsymbol{g}_{\boldsymbol{i}_c}\left(\boldsymbol{f}^{\boldsymbol{i}}(\boldsymbol{x})\right)\tag{17.29}$$

Then, by definition,

$$\int_A J^{*}(\boldsymbol{x})\,dx\equiv\int_A\det\left(D\boldsymbol{f}(\boldsymbol{x})D\boldsymbol{f}(\boldsymbol{x})^{*}\right)^{1/2}dx\tag{17.30}$$

First, using Theorem 17.5.7, and the fact that Lipschitz mappings take sets of measure zero to sets of measure zero, replace $E_j^{\boldsymbol{i}}$ with $\widetilde{E}_j^{\boldsymbol{i}}\subseteq E_j^{\boldsymbol{i}}$ such that $E_j^{\boldsymbol{i}}\setminus\widetilde{E}_j^{\boldsymbol{i}}$ has measure zero and

$$D\boldsymbol{f}^{\boldsymbol{i}}(\boldsymbol{g}(\boldsymbol{y}))D\boldsymbol{g}(\boldsymbol{y})=I,\ \left|\det(D\boldsymbol{g}(\boldsymbol{y}))\right|=\left|\det D\boldsymbol{f}^{\boldsymbol{i}}(\boldsymbol{g}(\boldsymbol{y}))\right|^{-1}\tag{17.31}$$

on $\boldsymbol{f}^{\boldsymbol{i}}\left(\widetilde{E}_j^{\boldsymbol{i}}\right)$. Changing the variables using the area formula and 17.31, the expression in 17.30 equals

$$\begin{aligned}\int_A J^{*}(\boldsymbol{x})\,dx &= \sum_{j=1}^{\infty}\int_{\widetilde{E}_j^{\boldsymbol{i}}\cap A}\sum_{\boldsymbol{i}\in\Lambda(n,m)}\left(\det\left(D\boldsymbol{f}^{\boldsymbol{i}}(\boldsymbol{x})D\boldsymbol{f}^{\boldsymbol{i}}(\boldsymbol{x})^{*}\right)\right)^{1/2}dx\\ &= \sum_{j=1}^{\infty}\sum_{\boldsymbol{i}\in\Lambda(n,m)}\int_{E_j^{\boldsymbol{i}}\cap A}\left(\det\left(D\boldsymbol{f}^{\boldsymbol{i}}(\boldsymbol{x})D\boldsymbol{f}^{\boldsymbol{i}}(\boldsymbol{x})^{*}\right)\right)^{1/2}dx\end{aligned}$$

$$=\sum_{j=1}^{\infty}\sum_{\boldsymbol{i}\in\Lambda(n,m)}\int_{\boldsymbol{f}^{\boldsymbol{i}}\left(E_j^{\boldsymbol{i}}\cap A\right)}\left(\det\left(D\boldsymbol{f}^{\boldsymbol{i}}(\boldsymbol{g}(\boldsymbol{y}))D\boldsymbol{f}^{\boldsymbol{i}}(\boldsymbol{g}(\boldsymbol{y}))^{*}\right)\right)^{1/2}\left|\det D\boldsymbol{f}^{\boldsymbol{i}}(\boldsymbol{g}(\boldsymbol{y}))\right|^{-1}dy\tag{17.32}$$

Recall everything is Borel measurable. I will consider one of the integrals in the sum.

$$\int_{E_j^i\cap A} \det\left(D\boldsymbol{f}^{\boldsymbol{i}}(\boldsymbol{x})D\boldsymbol{f}^{\boldsymbol{i}}(\boldsymbol{x})^*\right)^{1/2} dx \tag{17.33}$$

$$= \int_{\boldsymbol{f}^{\boldsymbol{i}}\left(E_j^i\cap A\right)} \det\left(D\boldsymbol{f}^{\boldsymbol{i}}(\boldsymbol{g}(\boldsymbol{y}))D\boldsymbol{f}^{\boldsymbol{i}}(\boldsymbol{g}(\boldsymbol{y}))^*\right)^{1/2} |D\boldsymbol{g}(\boldsymbol{y})|\, dy$$

$$= \int_{\boldsymbol{f}^{\boldsymbol{i}}\left(E_j^i\cap A\right)} \det\left(D\boldsymbol{f}^{\boldsymbol{i}}(\boldsymbol{g}(\boldsymbol{y}))D\boldsymbol{f}^{\boldsymbol{i}}(\boldsymbol{g}(\boldsymbol{y}))^*\right)^{1/2} \left|\det D\boldsymbol{f}^{\boldsymbol{i}}(\boldsymbol{g}(\boldsymbol{y}))\right|^{-1} dy$$

Now $\begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix} = \boldsymbol{f}^{\boldsymbol{i}}(\boldsymbol{x}) = \begin{pmatrix} \boldsymbol{f}(\boldsymbol{x}) \\ \boldsymbol{x}_{\boldsymbol{i}_c} \end{pmatrix}$ for $\boldsymbol{x}\in E_j^i\cap A$ if and only if $\boldsymbol{x}$ is also in $\boldsymbol{f}^{-1}(\boldsymbol{y}_1)$ which recall is a vector in $\mathbb{R}^n$. Therefore, by 17.27 and that $\boldsymbol{f}^{-1}(\boldsymbol{y}_1)\cap E_j^i\cap A$ is Borel measurable, the above equals the following iterated integral.

$$= \int_{\mathbb{R}^m}\int_{\boldsymbol{f}^{-1}(\boldsymbol{y}_1)\cap E_j^i\cap A} \det\left(D\boldsymbol{f}_{\boldsymbol{x}_{\boldsymbol{i}}}(\boldsymbol{g}(\boldsymbol{y}))D\boldsymbol{f}_{\boldsymbol{x}_{\boldsymbol{i}}}(\boldsymbol{g}(\boldsymbol{y}))^*\right)^{1/2} \left|\det D\boldsymbol{f}_{\boldsymbol{x}_{\boldsymbol{i}}}(\boldsymbol{g}(\boldsymbol{y}))\right|^{-1} dy_2 dy_1 \tag{17.34}$$

where $\boldsymbol{y}_1 = \boldsymbol{f}(\boldsymbol{x})$ and $\boldsymbol{y}_2 = \boldsymbol{x}_{\boldsymbol{i}_c}$. Since $\boldsymbol{y}_1$ is fixed in the inner integral of 17.34, and $\boldsymbol{y}_1 = \boldsymbol{f}(\boldsymbol{g}(\boldsymbol{y}))$, and by definition $\boldsymbol{g}_{\boldsymbol{i}_c}\left(\boldsymbol{f}^{\boldsymbol{i}}(\boldsymbol{x})\right) = \boldsymbol{y}_2$, one can take the partial derivative of $\boldsymbol{y}_1 = \boldsymbol{f}(\boldsymbol{g}(\boldsymbol{y}))$ with respect to $\boldsymbol{y}_2$ to obtain

$$\begin{aligned} \boldsymbol{0} &= D_{\boldsymbol{x}_{\boldsymbol{i}}}\boldsymbol{f}(\boldsymbol{g}(\boldsymbol{y}))D_{\boldsymbol{y}_2}\boldsymbol{g}_{\boldsymbol{i}}(\boldsymbol{y}) + D_{\boldsymbol{x}_{\boldsymbol{i}_c}}\boldsymbol{f}(\boldsymbol{g}(\boldsymbol{y}))D_{\boldsymbol{y}_2}\boldsymbol{g}_{\boldsymbol{i}_c}(\boldsymbol{y}) \\ &= D_{\boldsymbol{x}_{\boldsymbol{i}}}\boldsymbol{f}(\boldsymbol{g}(\boldsymbol{y}))D_{\boldsymbol{y}_2}\boldsymbol{g}_{\boldsymbol{i}}(\boldsymbol{y}) + D_{\boldsymbol{x}_{\boldsymbol{i}_c}}\boldsymbol{f}(\boldsymbol{g}(\boldsymbol{y})). \end{aligned} \tag{17.35}$$

Now consider the inner integral in 17.34 in which $\boldsymbol{y}_1$ is fixed. The integrand equals

$$\det\left[\left(D_{\boldsymbol{x}_{\boldsymbol{i}}}\boldsymbol{f}(\boldsymbol{g}(\boldsymbol{y}))\ D_{\boldsymbol{x}_{\boldsymbol{i}_c}}\boldsymbol{f}(\boldsymbol{g}(\boldsymbol{y}))\right)\begin{pmatrix} D_{\boldsymbol{x}_{\boldsymbol{i}}}\boldsymbol{f}(\boldsymbol{g}(\boldsymbol{y}))^* \\ D_{\boldsymbol{x}_{\boldsymbol{i}_c}}\boldsymbol{f}(\boldsymbol{g}(\boldsymbol{y}))^* \end{pmatrix}\right]^{1/2} \left|\det D\boldsymbol{f}_{\boldsymbol{x}_{\boldsymbol{i}}}(\boldsymbol{g}(\boldsymbol{y}))\right|^{-1}. \tag{17.36}$$

Let $A \equiv D_{\boldsymbol{x}_{\boldsymbol{i}}}\boldsymbol{f}(\boldsymbol{g}(\boldsymbol{y}))$ so A is $m\times m$ and $B \equiv D_{\boldsymbol{y}_2}\boldsymbol{g}_{\boldsymbol{i}}(\boldsymbol{y})$ an $m\times(n-m)$, and using 17.35, 17.36 is of the form

$$\begin{aligned} &\det\left[\begin{pmatrix} A & -AB \end{pmatrix}\begin{pmatrix} A^* \\ -B^*A^* \end{pmatrix}\right]^{1/2} |\det A|^{-1} \\ =\ &\det\left[AA^* + ABB^*A^*\right]^{1/2} |\det A|^{-1} \\ =\ &\det\left[A(I+BB^*)A^*\right]^{1/2} |\det A|^{-1} = \det(I+BB^*)^{1/2} \end{aligned}$$

which, by Theorem 17.8.1, equals $\det(I+B^*B)^{1/2}$. (Note the size of the identity changes in these two expressions.) Since $B = D_{\boldsymbol{y}_2}\boldsymbol{g}_{\boldsymbol{i}}(\boldsymbol{y})$ and $D_{\boldsymbol{y}_2}\boldsymbol{g}_{\boldsymbol{i}_c}(\boldsymbol{y}) = I$, the above reduces to

$$\det(I+B^*B)^{1/2} = \det\left[\begin{pmatrix} B^* & I \end{pmatrix}\begin{pmatrix} B \\ I \end{pmatrix}\right]^{1/2} =$$

$$\det\left[\begin{pmatrix} D_{\boldsymbol{y}_2}\boldsymbol{g}_{\boldsymbol{i}}(\boldsymbol{y})^* & D_{\boldsymbol{y}_2}\boldsymbol{g}_{\boldsymbol{i}_c}(\boldsymbol{y})^* \end{pmatrix}\begin{pmatrix} D_{\boldsymbol{y}_2}\boldsymbol{g}_{\boldsymbol{i}}(\boldsymbol{y}) \\ D_{\boldsymbol{y}_2}\boldsymbol{g}_{\boldsymbol{i}_c}(\boldsymbol{y}) \end{pmatrix}\right]^{1/2} = \det\left(D_{\boldsymbol{y}_2}\boldsymbol{g}(\boldsymbol{y})^* D_{\boldsymbol{y}_2}\boldsymbol{g}(\boldsymbol{y})\right)^{1/2}$$

Therefore, 17.33 reduces to $\int_{E_j^i\cap A}\det\left(D_{\boldsymbol{x}_i}\boldsymbol{f}(\boldsymbol{x})D_{\boldsymbol{x}_i}\boldsymbol{f}(\boldsymbol{x})^*\right)^{1/2}dx=$

$$\int_{\mathbb{R}^m}\int_{\boldsymbol{f}^{-1}(\boldsymbol{y}_1)\cap E_j^i\cap A}\det\left(D_{\boldsymbol{y}_2}\boldsymbol{g}(\boldsymbol{y})^*D_{\boldsymbol{y}_2}\boldsymbol{g}(\boldsymbol{y})\right)^{1/2}dy_2dy_1. \tag{17.37}$$

Then $\boldsymbol{z}\in\boldsymbol{g}\left(\boldsymbol{y}_1,\boldsymbol{f}^{-1}(\boldsymbol{y}_1)\cap E_j^i\cap A\right)$ if and only if

$$\boldsymbol{f}^{\boldsymbol{i}}(\boldsymbol{z})=\begin{pmatrix}\boldsymbol{f}(\boldsymbol{z})\\ \boldsymbol{z}_{i_c}\end{pmatrix}\in\begin{pmatrix}\boldsymbol{y}_1\\ \boldsymbol{f}^{-1}(\boldsymbol{y}_1)\cap E_j^i\cap A\end{pmatrix}$$

if and only if $\boldsymbol{z}\in\boldsymbol{f}^{-1}(\boldsymbol{y}_1)$ and $\boldsymbol{z}_{i_c}\in\boldsymbol{f}^{-1}(\boldsymbol{y}_1)\cap E_j^i\cap A$. Letting $\hat{\boldsymbol{g}}$ be the function $\boldsymbol{y}_2\to\boldsymbol{g}(\boldsymbol{y}_1,\boldsymbol{y}_2)$, this shows that $\boldsymbol{z}\in\hat{\boldsymbol{g}}\left(\boldsymbol{f}^{-1}(\boldsymbol{y}_1)\cap E_j^i\cap A\right)$ if and only if $\boldsymbol{y}_2=\boldsymbol{z}_{i_c}\in\boldsymbol{f}^{-1}(\boldsymbol{y}_1)\cap E_j^i\cap A$ and so $\hat{\boldsymbol{g}}\left(\boldsymbol{f}^{-1}(\boldsymbol{y}_1)\cap E_j^i\cap A\right)=\boldsymbol{f}^{-1}(\boldsymbol{y}_1)\cap E_j^i\cap A$. Of course $\hat{\boldsymbol{g}}$ actually depends on $\boldsymbol{y}_1$ but this is suppressed here. Therefore,

$$\boldsymbol{g}\left(\boldsymbol{y}_1,\boldsymbol{f}^{-1}(\boldsymbol{y}_1)\cap E_j^i\cap A\right)=\boldsymbol{f}^{-1}(\boldsymbol{y}_1)\cap E_j^i\cap A$$

By this observation and the area formula, the equations 17.33, 17.37 imply

$$\int_{E_j^i\cap A}\det\left(D\boldsymbol{f}^{\boldsymbol{i}}(\boldsymbol{x})D\boldsymbol{f}^{\boldsymbol{i}}(\boldsymbol{x})^*\right)^{1/2}dx=\int_{\mathbb{R}^m}\mathscr{H}^{n-m}\left(\boldsymbol{f}^{-1}(\boldsymbol{y}_1)\cap E_j^i\cap A\right)dy_1.$$

Therefore, summing the terms over all $\boldsymbol{i}$ and j,

$$\int_A\det\left(D\boldsymbol{f}(\boldsymbol{x})D\boldsymbol{f}(\boldsymbol{x})^*\right)^{1/2}dx=\int_{\mathbb{R}^m}\mathscr{H}^{n-m}\left(\boldsymbol{f}^{-1}(\boldsymbol{y})\cap A\right)dy$$

which verifies the coarea formula whenever A is a Borel set.

Borel measurability can be replaced with Lebesgue measurability. Let A be Lebesgue measurable and let $F\subseteq A\subseteq G$ where $m_n(G\setminus F)=0$ and F is F_σ and G is G_δ. See Theorem 9.4.6 along with the regularity of Lebesgue measure. Thus

$$\begin{aligned}&\int_{\mathbb{R}^m}\mathscr{H}^{n-m}\left(A\cap\boldsymbol{f}^{-1}(\boldsymbol{y})\right)dy\\ \in\ &\left[\int_{\mathbb{R}^m}\mathscr{H}^{n-m}\left(F\cap\boldsymbol{f}^{-1}(\boldsymbol{y})\right)dy,\int_{\mathbb{R}^m}\mathscr{H}^{n-m}\left(G\cap\boldsymbol{f}^{-1}(\boldsymbol{y})\right)dy\right]\\ =\ &\left[\int_F J^*(\boldsymbol{x})\,dx,\int_G J^*(\boldsymbol{x})\,dx\right]=\left[\int_A J^*(\boldsymbol{x})\,dx,\int_A J^*(\boldsymbol{x})\,dx\right]\ \blacksquare\end{aligned}$$

There is a simple corollary to this theorem in the case of locally Lipschitz maps.

Corollary 17.8.4 *Let $\boldsymbol{f}:\mathbb{R}^n\to\mathbb{R}^m$ where $m\le n$ and $\boldsymbol{f}$ is locally Lipschitz. This means that for each $r>0$, $\boldsymbol{f}$ is Lipschitz on $B(\boldsymbol{0},r)$. Then the coarea formula, 17.28, holds for $\boldsymbol{f}$.*

Proof: Let $A\subseteq B(\boldsymbol{0},r)$ and let $\boldsymbol{f}_r$ be Lipschitz with $\boldsymbol{f}(\boldsymbol{x})=\boldsymbol{f}_r(\boldsymbol{x})$ for $\boldsymbol{x}\in B(\boldsymbol{0},r+1)$. Then

$$\begin{aligned}&\int_A J^*\boldsymbol{f}(\boldsymbol{x})\,dx=\int_A J(D\boldsymbol{f}_r(\boldsymbol{x}))dx=\int_{\mathbb{R}^m}\mathscr{H}^{n-m}\left(A\cap\boldsymbol{f}_r^{-1}(\boldsymbol{y})\right)dy\\ =&\int_{\boldsymbol{f}_r(A)}\mathscr{H}^{n-m}\left(A\cap\boldsymbol{f}_r^{-1}(\boldsymbol{y})\right)dy=\int_{\boldsymbol{f}(A)}\mathscr{H}^{n-m}\left(A\cap\boldsymbol{f}^{-1}(\boldsymbol{y})\right)dy\end{aligned}$$

$$= \int_{\mathbb{R}^m} \mathscr{H}^{n-m}\left(A \cap \boldsymbol{f}^{-1}(\boldsymbol{y})\right) dy$$

Now for arbitrary Lebesgue measurable A the above shows for $k = 1, 2, \cdots$

$$\int_{A \cap B(\mathbf{0},k)} J^* \boldsymbol{f}(\boldsymbol{x})\, dx = \int_{\mathbb{R}^m} \mathscr{H}^{n-m}\left(A \cap B(\mathbf{0},k) \cap \boldsymbol{f}^{-1}(\boldsymbol{y})\right) dy.$$

Use the monotone convergence theorem to obtain 17.28. ■

Also note that by definition,

$$\int_{\mathbb{R}^m} \mathscr{H}^{n-m}\left(A \cap \boldsymbol{f}^{-1}(\boldsymbol{y})\right) dy = \int_{\boldsymbol{f}(A)} \mathscr{H}^{n-m}\left(A \cap \boldsymbol{f}^{-1}(\boldsymbol{y})\right) dy.$$

Recall that $\mathscr{H}^0(E)$ equals the number of elements in E. Thus, if $n = m$, the coarea formula implies

$$\int_A J^* \boldsymbol{f}(\boldsymbol{x})\, dx = \int_{\boldsymbol{f}(A)} \mathscr{H}^0\left(A \cap \boldsymbol{f}^{-1}(\boldsymbol{y})\right) dy = \int_{\boldsymbol{f}(A)} \#(\boldsymbol{y})\, dy \geq \int_{\boldsymbol{f}(A)} 1 dy$$

Thus, this gives a version of Sard's theorem by letting the singular set S be A in the above.

17.9 Change of Variables

Now let $s(\boldsymbol{x}) = \sum_{i=1}^p c_i \mathscr{X}_{E_i}(\boldsymbol{x})$ where E_i is Lebesgue measurable and $c_i \geq 0$. Then

$$\int_{\mathbb{R}^n} s(\boldsymbol{x}) J^* \boldsymbol{f}(\boldsymbol{x})\, dx = \sum_{i=1}^p c_i \int_{E_i} J^* \boldsymbol{f}(\boldsymbol{x})\, dx = \sum_{i=1}^p c_i \int_{\mathbb{R}^m} \mathscr{H}^{n-m}\left(E_i \cap \boldsymbol{f}^{-1}(\boldsymbol{y})\right) dy$$

$$= \int_{\mathbb{R}^m} \sum_{i=1}^p c_i \mathscr{H}^{n-m}\left(E_i \cap \boldsymbol{f}^{-1}(\boldsymbol{y})\right) dy = \int_{\mathbb{R}^m} \left[\int_{\boldsymbol{f}^{-1}(\boldsymbol{y})} s\, d\mathscr{H}^{n-m}\right] dy$$

$$= \int_{\mathbb{R}^m} \left[\int_{\boldsymbol{f}^{-1}(\boldsymbol{y})} s\, d\mathscr{H}^{n-m}\right] dy = \int_{\boldsymbol{f}(\mathbb{R}^n)} \left[\int_{\boldsymbol{f}^{-1}(\boldsymbol{y})} s\, d\mathscr{H}^{n-m}\right] dy. \tag{17.38}$$

Theorem 17.9.1 *Let $g \geq 0$ be Lebesgue measurable and let*

$$\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m,\ n \geq m,\ \boldsymbol{f} \text{ being Lipschitz}$$

Then

$$\int_{\mathbb{R}^n} g(\boldsymbol{x}) J^* \boldsymbol{f}(\boldsymbol{x})\, dx = \int_{\boldsymbol{f}(\mathbb{R}^n)} \left[\int_{\boldsymbol{f}^{-1}(\boldsymbol{y})} g(\boldsymbol{u})\, d\mathscr{H}^{n-m}(\boldsymbol{u})\right] dy. \tag{17.39}$$

Proof: Let $s_i \uparrow g$ where s_i is a simple function satisfying 17.38. Then let $i \to \infty$ and use the monotone convergence theorem to replace s_i with g. This proves the change of variables formula. ■

The following is an easy example of the use of the coarea formula to give a familiar relation.

Example 17.9.2 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be given by $f(\boldsymbol{x}) \equiv |\boldsymbol{x}|$. Then $J^*(\boldsymbol{x})$ ends up being 1. Then by the coarea formula,*

$$\int_{B(\mathbf{0},r)} dm_n = \int_0^r \mathscr{H}^{n-1}\left(B(\mathbf{0},r) \cap f^{-1}(y)\right) dy = \int_0^r \mathscr{H}^{n-1}(\partial B(\mathbf{0},y))\, dy$$

Then $m_n(B(\mathbf{0},r)) \equiv \alpha_n r^n = \int_0^r \mathscr{H}^{n-1}(\partial B(\mathbf{0},y))\,dy$. *Then differentiate both sides to obtain* $n\alpha_n r^{n-1} = \mathscr{H}^{n-1}(\partial B(\mathbf{0},r))$. *In particular* $\mathscr{H}^2(\partial B(\mathbf{0},r)) = 3\frac{4}{3}\pi r^2 = 4\pi r^2$. *Of course* α_n *was computed earlier. Recall from Theorem 14.5.1 on Page 411*

$$\alpha_n = \pi^{n/2}(\Gamma(n/2+1))^{-1}$$

Therefore, the $n-1$ *dimensional Hausdorf measure of the boundary of the ball of radius* r *in* $\mathbb{R}^n$ *is* $n\pi^{n/2}(\Gamma(n/2+1))^{-1}r^{n-1}$.

I think it is clear that you could generalize this to other more complicated situations. The above is nice because $J^*(\boldsymbol{x}) = 1$. This won't be so in general when considering other level surfaces.

17.10 Integration and the Degree

There is a very interesting application of the degree to integration [20]. I want to consider the case where $\boldsymbol{h}: \mathbb{R}^n \to \mathbb{R}^n$ is only Lipschitz continuous, vanishing outside a bounded set. In the following proposition, let ϕ_ε be a symmetric nonnegative mollifier,

$$\phi_\varepsilon(\boldsymbol{x}) \equiv \frac{1}{\varepsilon^n}\phi\left(\frac{\boldsymbol{x}}{\varepsilon}\right), \operatorname{spt}\phi \subseteq B(\mathbf{0},1).$$

Ω will be a bounded open set. By Rademacher's theorem, $\boldsymbol{h}$ satisfies

$$D\boldsymbol{h}(\boldsymbol{x}) \text{ exists } a.e., \tag{17.40}$$

If U is a bounded open set,

$$\lim_{m\to\infty} D(\boldsymbol{h}*\psi_m) = \lim_{m\to\infty} D\boldsymbol{h}*\psi_m = D\boldsymbol{h} \text{ in } L^1\left(U;\mathbb{R}^{n\times n}\right) \tag{17.41}$$

where ψ_m is a mollifier. Thus a subsequence converges a.e.

Now recall the definition of the degree.

Definition 17.10.1 *Let* Ω *be a bounded open set in* $\mathbb{R}^p$ *and let* $\boldsymbol{f}: \overline{\Omega} \to \mathbb{R}^p$ *be continuous. Let* $\boldsymbol{y} \notin \boldsymbol{f}(\partial\Omega)$. *Then the degree is defined as follows: Let* $\boldsymbol{g}$ *be infinitely differentiable,*

$$\|\boldsymbol{f}-\boldsymbol{g}\|_{\infty,\overline{\Omega}} < \delta \equiv \operatorname{dist}(\boldsymbol{f}(\partial\Omega),\boldsymbol{y}),$$

and $\boldsymbol{y}$ *is a regular value of* $\boldsymbol{g}$.*Then* $\boldsymbol{y} \notin \boldsymbol{g}(\partial\Omega)$ *and we define*

$$d(\boldsymbol{f},\Omega,\boldsymbol{y}) \equiv \sum\left\{\operatorname{sgn}(\det(D\boldsymbol{g}(\boldsymbol{x}))) : \boldsymbol{x} \in \boldsymbol{g}^{-1}(\boldsymbol{y}), \boldsymbol{x} \in \Omega\right\}$$

where the sum is finite by Lemma 15.1.5, defined to equal 0 if $\boldsymbol{g}^{-1}(\boldsymbol{y})$ *is empty.*

Also recall the fundamental integral identity.

Lemma 17.10.2 *Let* $\boldsymbol{y} \notin \boldsymbol{g}(\partial\Omega)$ *for* $\boldsymbol{g} \in C^\infty\left(\overline{\Omega};\mathbb{R}^p\right)$. *Also suppose* $\boldsymbol{y}$ *is a regular value of* $\boldsymbol{g}$. *Then for all positive* ε *small enough,*

$$\int_\Omega \phi_\varepsilon(\boldsymbol{g}(\boldsymbol{x})-\boldsymbol{y})\det D\boldsymbol{g}(\boldsymbol{x})\,dx = \sum\left\{\operatorname{sgn}(\det D\boldsymbol{g}(\boldsymbol{x})) : \boldsymbol{x} \in \boldsymbol{g}^{-1}(\boldsymbol{y})\right\}$$

The sum is the definition of the degree for $\boldsymbol{g}$ as described. There was also an important identity about homotopy Lemma 15.1.13 which includes the following.

Lemma 17.10.3 *If $\boldsymbol{h}$ is in $C^{\infty}\left(\overline{\Omega}\times[a,b],\mathbb{R}^{p}\right)$, and $\mathbf{0}\notin\boldsymbol{h}(\partial\Omega\times[a,b])$ then for $0<\varepsilon<\operatorname{dist}(\mathbf{0},\boldsymbol{h}(\partial\Omega\times[a,b]))$,*

$$t\to\int_{\Omega}\phi_{\varepsilon}(\boldsymbol{h}(\boldsymbol{x},t))\det D_{1}\boldsymbol{h}(\boldsymbol{x},t)\,dx$$

is constant for $t\in[a,b]$.

Let $\boldsymbol{h}_m$ in what follows be $\boldsymbol{h}*\psi_m$ for $\boldsymbol{h}$ continuous and ψ_m a mollifier. Eventually $\boldsymbol{h}$ will be Lipschitz continuous. Then $\boldsymbol{h}_m$ converges uniformly to $\boldsymbol{h}$ and so for all m large enough, $\|\boldsymbol{h}-\boldsymbol{h}_k\|_{\infty}<\delta\equiv\operatorname{dist}(\boldsymbol{y},\boldsymbol{h}(\partial\Omega))$, say for all $k\geq m$, and so for a fixed $\boldsymbol{y}\notin\boldsymbol{h}(\partial\Omega)$ and homotopy invariance of the degree,

$$d(\boldsymbol{h},\Omega,\boldsymbol{y})=d(\boldsymbol{h}_m,\Omega,\boldsymbol{y})=d(\boldsymbol{h}_m,\Omega,\tilde{\boldsymbol{y}})$$

where $\tilde{\boldsymbol{y}}$ is a regular value of $\boldsymbol{h}_m$ which is close to $\boldsymbol{y}$ and contained $\boldsymbol{h}(\partial\Omega)^{C}$. We can also take $\tilde{\boldsymbol{y}}$ so close to $\boldsymbol{y}$ that $\boldsymbol{h}_m(\boldsymbol{x})-(\boldsymbol{y}+t(\tilde{\boldsymbol{y}}-\boldsymbol{y}))\neq\mathbf{0}$ for all $t\in[0,1],\boldsymbol{x}\in\partial\Omega$. Therefore, there is $\varepsilon>0$ such that for $\boldsymbol{h}(\boldsymbol{x},t)\equiv\boldsymbol{h}_m(\boldsymbol{x})-(\boldsymbol{y}+t(\tilde{\boldsymbol{y}}-\boldsymbol{y})),0<\varepsilon<\boldsymbol{h}(\partial\Omega\times[0,1])$. Thus the following is constant for $t\in[0,1]$

$$\int_{\Omega}\phi_{\varepsilon}(\boldsymbol{h}_m(\boldsymbol{x})-(\boldsymbol{y}+t(\tilde{\boldsymbol{y}}-\boldsymbol{y})))\det(D\boldsymbol{h}_m(\boldsymbol{x}))\,dx$$

Therefore, making ε still smaller if necessary, so that the integral relation for the degree holds,

$$\begin{aligned}
d(\boldsymbol{h}_m,\Omega,\tilde{\boldsymbol{y}}) &= \int_{\Omega}\phi_{\varepsilon}(\boldsymbol{h}_m(\boldsymbol{x})-\tilde{\boldsymbol{y}})\det(D\boldsymbol{h}_m(\boldsymbol{x}))\,dx\\
&= \int_{\Omega}\phi_{\varepsilon}(\boldsymbol{h}_m(\boldsymbol{x})-(\boldsymbol{y}+1(\tilde{\boldsymbol{y}}-\boldsymbol{y})))\det(D\boldsymbol{h}_m(\boldsymbol{x}))\,dx\\
&= \int_{\Omega}\phi_{\varepsilon}(\boldsymbol{h}_m(\boldsymbol{x})-(\boldsymbol{y}+0(\tilde{\boldsymbol{y}}-\boldsymbol{y})))\det(D\boldsymbol{h}_m(\boldsymbol{x}))\,dx\\
&= \int_{\Omega}\phi_{\varepsilon}(\boldsymbol{h}_m(\boldsymbol{x})-\boldsymbol{y})\det(D\boldsymbol{h}_m(\boldsymbol{x}))\,dx=d(\boldsymbol{h}_m,\Omega,\boldsymbol{y})
\end{aligned}$$

This, along with properties of he degree yields the following interesting result.

Lemma 17.10.4 *Let $\boldsymbol{h}\in C\left(\overline{\Omega};\mathbb{R}^{n}\right)$ and let $\boldsymbol{y}\notin\boldsymbol{h}(\partial\Omega),\delta\equiv\operatorname{dist}(\boldsymbol{y},\boldsymbol{h}(\partial\Omega))$. Let $\boldsymbol{h}_m$ be in $C^{\infty}\left(\overline{\Omega};\mathbb{R}^{n}\right)$ and $\|\boldsymbol{h}_m-\boldsymbol{h}\|_{\infty}<\delta$. Then by homotopy invariance in the first equation and what was just shown,*

$$d(\boldsymbol{h},\Omega,\boldsymbol{y})=d(\boldsymbol{h}_m,\Omega,\boldsymbol{y})=\int_{\Omega}\phi_{\varepsilon}(\boldsymbol{h}_m(\boldsymbol{x})-\boldsymbol{y})\det(D\boldsymbol{h}_m(\boldsymbol{x}))\,dx$$

for some ε sufficiently small.

Let ε be smaller if necessary such that the above holds and also

$$0<\varepsilon<\operatorname{dist}(\boldsymbol{y},\boldsymbol{h}(\partial\Omega)).$$

Then for $\hat{\boldsymbol{y}}$ close to $\boldsymbol{y}$,

$$\int_\Omega \phi_\varepsilon(\boldsymbol{h}_m(\boldsymbol{x}) - (\boldsymbol{y} + t(\hat{\boldsymbol{y}} - \boldsymbol{y}))) \det(D\boldsymbol{h}_m(\boldsymbol{x}))\, dx$$

is constant for $t \in [0,1]$ thanks to Lemma 17.10.3, and

$$d(\boldsymbol{h}_m, \Omega, \boldsymbol{y}) = d(\boldsymbol{h}_m, \Omega, \hat{\boldsymbol{y}}) = \int_\Omega \phi_\varepsilon(\boldsymbol{h}_m(\boldsymbol{x}) - \hat{\boldsymbol{y}}) \det(D\boldsymbol{h}_m(\boldsymbol{x}))\, dx.$$

Say this is true for $\hat{\boldsymbol{y}} \in B(\boldsymbol{y}, r)$. Now let $f \in C_c(B(\boldsymbol{y}, r))$.

$$\begin{aligned}\int f(\boldsymbol{z})\, d(\boldsymbol{h}_m, \Omega, \boldsymbol{z})\, dz &= \int f(\boldsymbol{z}) \int_\Omega \phi_\varepsilon(\boldsymbol{h}_m(\boldsymbol{x}) - \boldsymbol{z}) \det(D\boldsymbol{h}_m(\boldsymbol{x}))\, dx dz \\ &= \int_\Omega \det(D\boldsymbol{h}_m(\boldsymbol{x})) \int f(\boldsymbol{z})\, \phi_\varepsilon(\boldsymbol{h}_m(\boldsymbol{x}) - \boldsymbol{z})\, dz dx\end{aligned}$$

Now this holds for all ε sufficiently small and so we can pass to a limit and use uniform continuity of f to conclude that

$$\int f(\boldsymbol{z})\, d(\boldsymbol{h}_m, \Omega, \boldsymbol{z})\, dz = \int_\Omega f(\boldsymbol{h}_m(\boldsymbol{x})) \det(D\boldsymbol{h}_m(\boldsymbol{x}))\, dx$$

Next suppose $f \in C_c\left(\boldsymbol{h}(\partial\Omega)^C\right)$. Then $\operatorname{spt}(f)$ is covered by finitely many of such balls like the above, $\{B_i\}_{i=1}^m$ with the property that if $g \in C_c(B_i)$, then

$$\int g(\boldsymbol{z})\, d(\boldsymbol{h}_m, \Omega, \boldsymbol{z})\, dz = \int_\Omega g(\boldsymbol{h}_m(\boldsymbol{x})) \det(D\boldsymbol{h}_m(\boldsymbol{x}))\, dx$$

Now let $\left\{\psi_j\right\}$ be a partition of unity on $\operatorname{spt} f$ with $\operatorname{spt}\psi_i \subseteq B_i$. Then

$$\int f(\boldsymbol{z})\, d(\boldsymbol{h}_m, \Omega, \boldsymbol{z})\, dz = \int \sum_{i=1}^m \psi_i(\boldsymbol{z}) f(\boldsymbol{z})\, d(\boldsymbol{h}_m, \Omega, \boldsymbol{z})\, dz$$

$$= \sum_{i=1}^m \int \psi_i(\boldsymbol{z}) f(\boldsymbol{z})\, d(\boldsymbol{h}_m, \Omega, \boldsymbol{z})\, dz = \sum_{i=1}^m \int_\Omega \psi_i(\boldsymbol{h}_m(\boldsymbol{x})) f(\boldsymbol{h}_m(\boldsymbol{x})) \det(D\boldsymbol{h}_m(\boldsymbol{x}))\, dx$$

$$= \int_\Omega \sum_{i=1}^m \psi_i(\boldsymbol{h}_m(\boldsymbol{x})) f(\boldsymbol{h}_m(\boldsymbol{x})) \det(D\boldsymbol{h}_m(\boldsymbol{x}))\, dx = \int_\Omega f(\boldsymbol{h}_m(\boldsymbol{x})) \det(D\boldsymbol{h}_m(\boldsymbol{x}))\, dx$$

If $\boldsymbol{h}$ is Lipschitz, then $\lim_{m\to\infty} \boldsymbol{h}_m(\boldsymbol{x}) = \boldsymbol{h}(\boldsymbol{x})$ uniformly and also for a.e. $\boldsymbol{x}$,

$$D\boldsymbol{h}_m(\boldsymbol{x}) \to D\boldsymbol{h}(\boldsymbol{x})$$

and so, since everything is bounded, we can apply the dominated convergence theorem and conclude that

$$\int f(\boldsymbol{z})\, d(\boldsymbol{h}, \Omega, \boldsymbol{z})\, dz = \int_\Omega f(\boldsymbol{h}(\boldsymbol{x})) \det(D\boldsymbol{h}(\boldsymbol{x}))\, dx$$

This proves the following lemma.

Lemma 17.10.5 *Let $f \in C_c\left(\boldsymbol{h}(\partial\Omega)^C\right)$ and let $\boldsymbol{h}$ be Lipschitz on $\mathbb{R}^n$. Then*

$$\int f(\boldsymbol{y})\, d(\boldsymbol{h}, \Omega, \boldsymbol{y})\, dy = \int_\Omega f(\boldsymbol{h}(\boldsymbol{x})) \det(D\boldsymbol{h}(\boldsymbol{x}))\, dx.$$

I will stop at this point. However, you could probably say more. For example, the degree is constant on components of $\boldsymbol{h}(d\Omega)^C$ and so considering these components, you could use the Riesz representation theorem for positive linear functionals to get this formula for more general f.

Chapter 18

Differential Forms

This is a generalization of single variable ideas from calculus to multiple dimensions. It generalizes the fundamental theorem of calculus of the form $\int_a^b f'(t)\,dt = f(b) - f(a)$ and integration by parts. The challenge is in keeping track of orientation. This is why one considers differential forms and much linear algebra machinery. Most of what follows involves whatever assumptions of smoothness are convenient, but to extend to the case of Lipschitz mappings, I will use Lemma 16.4.1.

Also note that the composition of Lipschitz mappings is Lipschitz and the usual chain rule will hold off some set of measure zero.

Recall from calculus that when you had a differential form, written as

$$\int_C a(x,y,z)\,dx + b(x,y,z)\,dy + c(x,y,z)\,dz$$

where C is an oriented curve, it gave the work done by the force field

$$\boldsymbol{F}(x,y,z) = (a(x,y,z), b(x,y,z), c(x,y,z))$$

on an object moving over the oriented curve C. How was it evaluated? You had a parametrization

$$\boldsymbol{r} : [a,b] \to \mathbb{R}^3, \boldsymbol{r}(t) = (x(t), y(t), z(t))$$

and then you did the following.

$$\int_a^b \left(a(x(t),y(t),z(t)) \frac{dx}{dt} + b(x(t),y(t),z(t)) \frac{dy}{dt} + c(x(t),y(t),z(t)) \frac{dz}{dt} \right) dt$$

Note how the orientation of the curve comes from the interval and the choice of parameterization. The interval $[a,b]$ is called the parameter domain and t is referred to as a parameter. The curve itself is some object in $\mathbb{R}^3$.

You can think of the differential form $a(x,y,z)\,dx + b(x,y,z)\,dy + c(x,y,z)\,dz$ as a symbol which represents something which makes vector valued functions $\boldsymbol{r}$ defined on $[a,b]$ into numbers according to the above procedure. You might also have written the line integral in the form

$$\int_{\boldsymbol{r}} a(x,y,z)\,dx + b(x,y,z)\,dy + c(x,y,z)\,dz$$

Thus it is a functional which makes vector valued functions defined on $[a,b]$ into numbers. It is desired to extend this simple idea to functions of $k > 1$ variables. That which will take the place of an interval will be a box $\prod_{j=1}^k [a_j, b_j]$ or a chain of boxes formed by pasting boxes together along a common face. The latter will give a more general parameter domain than just a box. This is analogous to the notion of piecewise smooth curves where the curve was obtained by joining one smooth curve to another at an end point.

First here is some notation.

Notation 18.0.1 *Let $\boldsymbol{r} : \prod_{j=1}^k [a_j, b_j] \to \mathbb{R}^p$, $p \geq k$ with $\boldsymbol{r} \in C^1\left(\prod_{j=1}^k [a_j, b_j], \mathbb{R}^p\right)$. This last symbol means that $\boldsymbol{r}$ is the restriction to $\prod_{j=1}^k [a_j, b_j]$ of a C^1 function defined on all of $\mathbb{R}^k$. Now let I denote an ordered list of k indices taken from $\{1, 2, \cdots, p\}$. Thus*

$I = (i_1, \cdots, i_k)$. Then

$$\det\left(\frac{d\boldsymbol{r}^I}{d\boldsymbol{u}}\right) \equiv \det\begin{pmatrix} x_{i_1,u_1} & x_{i_1,u_2} & \cdots & x_{i_1,u_k} \\ x_{i_2,u_1} & x_{i_2,u_2} & \cdots & x_{i_2,u_k} \\ \vdots & \vdots & & \vdots \\ x_{i_k,u_1} & x_{i_k,u_2} & \cdots & x_{i_k,u_k} \end{pmatrix} \equiv \frac{\partial\left(x_{i_1}, \cdots x_{i_k}\right)}{\partial\left(u_1, \cdots, u_k\right)}$$

It is the same as $\det\left(D\boldsymbol{r}^I\right)$ *where* $\boldsymbol{r}^I$ *has values in* $\mathbb{R}^p$ *and is obtained by keeping the rows of* $\boldsymbol{r}$ *in the order determined by* I *and leaving out the other rows. More generally, suppose* I *is an ordered list of* l *indices and that* J *is an ordered list of* l *indices. Then*

$$\det\left(\frac{d\boldsymbol{r}^I}{d\boldsymbol{u}_J}\right) \equiv \det\begin{pmatrix} x_{i_1,u_{j_1}} & x_{i_1,u_{j_2}} & \cdots & x_{i_1,u_{j_l}} \\ x_{i_2,u_{j_1}} & x_{i_2,u_{j_2}} & \cdots & x_{i_2,u_{j_l}} \\ \vdots & \vdots & & \vdots \\ x_{i_l,u_{j_1}} & x_{i_l,u_{j_2}} & \cdots & x_{i_l,u_{j_l}} \end{pmatrix} \equiv \frac{\partial\left(x_{i_1}, \cdots x_{i_l}\right)}{\partial\left(u_{j_1}, \cdots, u_{j_l}\right)},\ \boldsymbol{x} = \boldsymbol{r}(\boldsymbol{u})$$

Now with this definition, here is the generalization of the differential forms defined in calculus.

Definition 18.0.2 *A differential form is* $\omega \equiv \sum_I a_I(\boldsymbol{x})\, d\boldsymbol{x}^I$ *where*

$$d\boldsymbol{x}^I \equiv dx_{i_1} \wedge dx_{i_2} \wedge \cdots \wedge dx_{i_k}$$

To save space, let $[\boldsymbol{a},\boldsymbol{b}] \equiv \prod_{k=1}^{k} [a_j, b_j]$. *For* $I = (i_1, \cdots, i_k)$, $\int_{(\cdot)} \omega$ *is a function mapping functions* $\boldsymbol{r}$ *in* $C^1\left(\prod_{j=1}^{k} [a_j, b_j], \mathbb{R}^p\right)$ *or Lipschitz functions to* $\mathbb{R}$ *, defined by*

$$\begin{aligned} \int_{\boldsymbol{r}} \omega &\equiv \int_{[\boldsymbol{a},\boldsymbol{b}]} \sum_I a_I(\boldsymbol{r}(\boldsymbol{u})) \det\left(\frac{d\boldsymbol{x}^I(\boldsymbol{u})}{d\boldsymbol{u}}\right) dm_k \\ &= \int_{[\boldsymbol{a},\boldsymbol{b}]} \sum_I a_I(\boldsymbol{r}(\boldsymbol{u})) \det\left(\frac{d\boldsymbol{r}^I(\boldsymbol{u})}{d\boldsymbol{u}}\right) dm_k \end{aligned}$$

The sum is taken over all ordered lists of indices from $\{1, \cdots, p\}$. *Note that if there are any repeats in an ordered list* I, *then* $\det\left(\frac{d\boldsymbol{r}^I(\boldsymbol{u})}{d\boldsymbol{u}}\right) = 0$ *and so it suffices to consider the sum only over lists of indices in which there are no repeats. Thus the sum can be considered to consist of no more than* $P(p,k)$ *terms where this denotes the permutations of* p *things taken* k *at a time.*

Actually, if I, J are the same set of indices, listed in different order, then $\det\left(\frac{d\boldsymbol{x}^I(\boldsymbol{u})}{d\boldsymbol{u}}\right)$ will be $\pm\det\left(\frac{d\boldsymbol{x}^J(\boldsymbol{u})}{d\boldsymbol{u}}\right)$ so you could always write the differential form in terms of sums over strictly increasing lists of indices, and if this is done, you can always have only $C(p,k) = \frac{p!}{k!(p-k)!}$ terms in the sum where this denotes combinations of p things taken k at a time.

It will always be assumed that a_I is as smooth as desired to make everything work. As to $\boldsymbol{r} \in C^1([\boldsymbol{a},\boldsymbol{b}], \mathbb{R}^p)$ or Lipschitz, it is not required to have $D\boldsymbol{r}^I(\boldsymbol{u})$ be nonzero for any I. This means $\boldsymbol{r}([\boldsymbol{a},\boldsymbol{b}])$ can be various sets which have points and edges.

Note that if $p < k$, then the functional $\sum_I a_I(\boldsymbol{x})\, d\boldsymbol{x}^I$ should equal the zero function because you would have $x_{p+1} = \cdots = x_k = 0$ and so there would be at least one row of zeros in the above determinant. Thus, I will suppose that $k \leq p$ in what follows.

Example 18.0.3 *Consider the ball $B(\mathbf{0},r)$. Spherical coordinates are $\mathbf{r}=(x,y,z)$ where*

$$x=\rho\sin(\phi)\cos(\theta),\ y=\rho\sin(\phi)\sin(\theta),z=\rho\cos(\phi)$$

Let $(\rho,\phi,\theta)\in[0,r]\times[0,\pi]\times[0,2\pi]$ which is a box like what was just described. This mapping is C^1 and onto the ball. However, it is clearly not one to one. However, $\mathbf{r}$ is one to one on $(0,r]\times[0,\pi]\times[0,2\pi)$ off a closed set S of m_3 measure zero. Also $\mathbf{r}(S)$ has measure zero by Sard's theorem. Recall from calculus that $\frac{\partial(x,y,z)}{\partial(\rho,\phi,\theta)}=\rho^2\sin\phi$ which is positive if $\rho>0$ and $\theta\in(0,\pi)$.

In a similar manner, you could obtain a solid ellipse.

$$x=a\sin(\phi)\cos(\theta),\ y=b\sin(\phi)\sin(\theta),z=c\cos(\phi)$$

for $a,b,c>0$.

Example 18.0.4 *Consider the boundary of the ball $B(\mathbf{0},r)$. Spherical coordinates are*

$$x=r\sin(\phi)\cos(\theta),\ y=r\sin(\phi)\sin(\theta),z=r\cos(\phi)$$

Let $(\phi,\theta)\in[0,\pi]\times[0,2\pi]$ which is a box like what was just described. This mapping is C^1 and onto the boundary of this ball. However, it is clearly not one to one. However, off a set S of m_2 measure zero, the mapping is indeed one to one on what remains of this box and $\mathbf{r}(S)$ has $\mathscr{H}^2$ measure zero.

$$\frac{\partial(x,y)}{\partial(\phi,\theta)}=\begin{vmatrix} r\cos\phi\cos\theta & -r\sin\phi\sin\theta \\ r\cos\phi\sin\theta & r\sin\phi\cos\theta \end{vmatrix}=r^2\cos\phi\sin\phi$$

$$\frac{\partial(x,z)}{\partial(\phi,\theta)}=\begin{vmatrix} r\cos\phi\cos\theta & -r\sin\phi \\ r\cos\phi\sin\theta & 0 \end{vmatrix}=r^2\cos(\phi)\sin(\phi)\sin(\theta)$$

$$\frac{\partial(y,z)}{\partial(\phi,\theta)}=\begin{vmatrix} -r\sin\phi\sin\theta & -r\sin\phi \\ r\sin\phi\cos\theta & 0 \end{vmatrix}=r^2\sin^2(\phi)\cos(\theta)$$

Similarly, you could obtain the boundary of an ellipse in the same way.

Example 18.0.5 *One can obtain the cylinder of radius r of height h as follows.*

$$x=r\cos(\theta),y=r\sin(\theta),z$$

where $(\theta,z)\in[0,2\pi]\times[0,h]$ a box. The mapping is clearly C^1 but is not one to one.

More generally,

Example 18.0.6 *Say $u\to(a(u),b(u))$ for $u\in[a,b]$ is a curve in the plane where a,b are C^1 and you consider*

$$x=a(u),y=b(u),z=s$$

where $(s,u)\in[c,d]\times[a,b]$. This would be a surface in three dimensions. It may or may not be a one to one mapping depending on whether the curve is or is not a closed curve. It would not be one to one if the curve is a closed curve with $(x(b),y(b))=(x(a),y(a))$.

Many other examples are available including higher dimensional versions of the above.

Lemma 18.0.7 *Each differential form ω can be written in a unique way as*

$$\sum_{i_1<i_2<\cdots<i_k} a_{i_1,i_2,\cdots,i_k}(\boldsymbol{x})\, dx_{i_1}\wedge dx_{i_2}\wedge\cdots\wedge dx_{i_k}$$

Also if σ is a permutation and $J=\sigma(I)$, then the following holds

$$d\boldsymbol{x}^J = \operatorname{sgn}(\sigma)\, d\boldsymbol{x}^I$$

Proof: Suppose you have a differential form $\omega=\sum_I a_I(\boldsymbol{x})\, d\boldsymbol{x}^I$. Then for a given $\boldsymbol{r}\in C^1([\boldsymbol{a},\boldsymbol{b}],\mathbb{R}^p)$ and $\{I\}=\{J\}$ meaning that

$$\{i_1,\cdots,i_k\}=\{j_1,\cdots,j_k\}$$

so that J is a permutation of I, it follows that

$$\int_{\boldsymbol{r}} a_J(\boldsymbol{r}(\boldsymbol{u}))\det\left(\frac{d\boldsymbol{r}_J(\boldsymbol{u})}{d\boldsymbol{u}}\right)dm_k = \operatorname{sgn}(\sigma)\int_{\boldsymbol{r}} a_J(\boldsymbol{r}(\boldsymbol{u}))\det\left(\frac{d\boldsymbol{r}^I(\boldsymbol{u})}{d\boldsymbol{u}}\right)dm_k$$

where $\operatorname{sgn}(\sigma)$ is the sign of the permutation which takes I to J. This follows from the alternating property of determinants. This verifies the second claim. Thus, considering the two terms

$$a_I(\boldsymbol{x})\, d\boldsymbol{x}^I, a_J(\boldsymbol{x})\, d\boldsymbol{x}^J$$

in the sum defining the differential form, the effect on $\boldsymbol{r}$ would be the same as that of the single term

$$(a_I(\boldsymbol{x})+a_J(\boldsymbol{x})\operatorname{sgn}(\sigma))\, d\boldsymbol{x}^I$$

combining these in this way leads to a sum of the form

$$\omega=\sum_{i_1<i_2<\cdots<i_k} a_{i_1,i_2,\cdots,i_k}(\boldsymbol{x})\, dx_{i_1}\wedge dx_{i_2}\wedge\cdots\wedge dx_{i_k} \tag{18.1}$$

where the $a_{i_1,i_2,\cdots,i_k}$ are obtained from the original a_I by combining them as just explained.

Now it will be shown that the above sum in 18.1 is uniquely determined. This can be shown by considering a particular function in $C^1([\boldsymbol{a},\boldsymbol{b}],\mathbb{R}^p)$ which is chosen auspiciously to reveal $a_{i_1,i_2,\cdots,i_k}(\boldsymbol{x}), \boldsymbol{x}\in(\boldsymbol{a},\boldsymbol{b})$. Letting I be an ordered list of k indices from $\{1,\cdots,p\}$. Say $I=(i_1,\cdots,i_k)$ with $i_1<i_2<\cdots<i_k$, an ascending list, meaning that the indices are increasing. Define $\boldsymbol{r}_\delta(\boldsymbol{u})$ such that if $j\notin\{i_1,\cdots,i_k\}, r_{\delta j}(u)\equiv x_j$ and if $j=i_s, s=1,2,...,k$, then $r_{\delta j}(u)\equiv x_j+\frac{\delta}{(b_s-a_s)}(u_s-a_j)$. Thus $\boldsymbol{r}_\delta(\boldsymbol{u})\equiv$

$$\begin{array}{rl} & \left(\begin{array}{ccccccc} x_1 & \cdots & x_{i_1}+\frac{\delta}{b_1-a_1}(u_1-a_1) & \cdots & x_{i_k}+\frac{\delta}{b_k-a_k}(u_k-a_k) & \cdots & x_p \end{array}\right)^T \\ \equiv & \boldsymbol{x}+\boldsymbol{r}_\delta(\boldsymbol{u}) \end{array}$$

So what is $\int_{\boldsymbol{r}_\delta}\omega$? By definition, it equals

$$\begin{aligned} \int_{\boldsymbol{r}_\delta}\omega &= \sum_{J\text{ ascending}}\int_{[\boldsymbol{a},\boldsymbol{b}]} a_J(\boldsymbol{r}_\delta(\boldsymbol{u}))\det\left(\frac{d\boldsymbol{r}_{\delta J}(\boldsymbol{u})}{d\boldsymbol{u}}\right)dm_k \\ &= \prod_{j=1}^k\left(\frac{\delta}{b_j-a_j}\right)\int_{[\boldsymbol{a},\boldsymbol{b}]} a_I(\boldsymbol{x}+\boldsymbol{d}_\delta(\boldsymbol{u}))\, dm_k(\boldsymbol{u}) \end{aligned} \tag{18.2}$$

because if J does not consist of the same indices as I, then there must be a row of zeros in $\frac{d\boldsymbol{r}_{\delta J}(\boldsymbol{u})}{d\boldsymbol{u}}$ and so the integral for that term in the above formula must equal 0. Let $(u_j - a_j)\frac{\delta}{b_j-a_j} = u_j,\ j = 1,...,k$ so $\boldsymbol{u} \in [\boldsymbol{0},\boldsymbol{\delta}] \equiv \prod_{s=1}^{k}[0,\delta]$. The appropriate Jacobian is then $\left(\prod_{j=1}^{k}\left(\frac{\delta}{b_j-a_j}\right)\right)^{-1}$ and so 18.2 reduces to $\int_{[\boldsymbol{0},\boldsymbol{\delta}]} a_I(\boldsymbol{x}+\boldsymbol{u}\boldsymbol{e}_I)\,dm_k(u)$. Thus

$$\delta^{-k}\int_{\boldsymbol{r}_\delta}\boldsymbol{\omega} = \delta^{-k}\int_{[\boldsymbol{0},\boldsymbol{\delta}]} a_I(\boldsymbol{x}+\boldsymbol{u}\boldsymbol{e}_I)\,dm_k(u)$$

so by the fundamental theorem of calculus, $\lim_{\delta\to 0}\delta^{-k}\int_{\boldsymbol{r}_\delta}\boldsymbol{\omega} = a_I(\boldsymbol{x})$. This verifies uniqueness. ■

18.1 The Wedge Product

Next is the definition of the wedge product.

Definition 18.1.1 *Denote differential forms, the functionals defined above which act on $C^1([\boldsymbol{a},\boldsymbol{b}],\mathbb{R}^p)$ as Ω^k. Then letting $\boldsymbol{\omega}\in\Omega^k$ and $\eta\in\Omega^l$, let $\boldsymbol{\omega}\wedge\eta\in\Omega^{k+l}$ be defined as follows. Letting*

$$\boldsymbol{\omega} \equiv \sum_{I \text{ ascending}} a_I(\boldsymbol{x})\,d\boldsymbol{x}^I,\quad \eta \equiv \sum_{J \text{ ascending}} b_J(\boldsymbol{x})\,d\boldsymbol{x}^J$$

then

$$\boldsymbol{\omega}\wedge\eta \equiv \sum_{I \text{ ascending}}\sum_{J \text{ ascending}} a_I(\boldsymbol{x})\,b_J(\boldsymbol{x})\,d\boldsymbol{x}^I\wedge d\boldsymbol{x}^J$$

where for $I=(i_1,\cdots,i_k), J\equiv(j_1,\cdots,j_l)$,

$$d\boldsymbol{x}^I\wedge d\boldsymbol{x}^J \equiv dx_{i_1}\wedge\cdots\wedge dx_{i_k}\wedge dx_{j_1}\wedge\cdots\wedge dx_{j_l}$$

This is well defined thanks to the above lemma which shows that there is only one way to write a differential form like the above sums in which I is ascending in each term.

What if I and J are not ascending? Does it still work? Let $I=(i_1,\cdots,i_k), J\equiv(j_1,\cdots,j_l)$ and let $\sigma(I)$ be ascending and let $\eta(J)$ be ascending, σ,η being two permutations. Then from Lemma 18.0.7, the second claim,

$$\begin{aligned}
\overset{\in\Omega^k}{d\boldsymbol{x}^I}\wedge\overset{\in\Omega^l}{d\boldsymbol{x}^J} &= \overset{\text{writing each in ascending form}}{\operatorname{sgn}(\sigma)\,d\boldsymbol{x}_{\sigma(I)}\wedge\operatorname{sgn}(\eta)\,d\boldsymbol{x}_{\eta(J)}}\\
&\overset{\text{above definition}}{=} \operatorname{sgn}(\sigma)\operatorname{sgn}(\eta)\,d\boldsymbol{x}_{\sigma(I)}\wedge d\boldsymbol{x}_{\eta(J)}\\
&\overset{\text{definition of the wedge product } d\boldsymbol{x}_{\sigma(I)}\wedge d\boldsymbol{x}_{\eta(J)}}{=} \operatorname{sgn}(\sigma)\operatorname{sgn}(\eta)\,dx_{\sigma(i_1)}\wedge\cdots\wedge dx_{\sigma(i_k)}\wedge dx_{\eta(j_1)}\wedge\cdots\wedge dx_{\eta(j_l)}\\
&= dx_{i_1}\wedge\cdots\wedge dx_{i_k}\wedge dx_{j_1}\wedge\cdots\wedge dx_{j_l}
\end{aligned}$$

Thus it is correct to write

$$d\boldsymbol{x}^I\wedge d\boldsymbol{x}^J = dx_{i_1}\wedge\cdots\wedge dx_{i_k}\wedge dx_{j_1}\wedge\cdots\wedge dx_{j_l}$$

even if $I=(i_1,\cdots,i_k), J\equiv(j_1,\cdots,j_l)$ are not ascending.

This is the main idea behind the following fundamental result. In this lemma, I,J are not necessarily ascending.

Lemma 18.1.2 *Let $\omega \equiv \sum_I a_I(\boldsymbol{x})\, d\boldsymbol{x}^I, \eta \equiv \sum_J b_J(\boldsymbol{x})\, d\boldsymbol{x}^J$ be in Ω^k and Ω^l respectively. Then*

$$\omega \wedge \eta = \sum_{I,J} a_I(\boldsymbol{x})\, b_J(\boldsymbol{x})\, d\boldsymbol{x}^I \wedge d\boldsymbol{x}^J$$

Proof: For each I, let $\sigma_I(I)$ be ascending where σ_I is a permutation, similar for J. Then, as noted above in the proof of Lemma 18.0.7 and denoting by $\hat{I}, \hat{J}$ the ascending lists of indices.

$$\begin{aligned}
\omega &= \sum_{\hat{I}} \left(\sum_{\{I:\sigma_I(I)=\hat{I}\}} a_I(\boldsymbol{x}) \operatorname{sgn}(\sigma_I) \right) d\boldsymbol{x}^{\hat{I}} \\
\eta &= \sum_{\hat{J}} \left(\sum_{\{I:\sigma_J(J)=\hat{J}\}} b_J(\boldsymbol{x}) \operatorname{sgn}(\sigma_J) \right) d\boldsymbol{x}^{\hat{J}}
\end{aligned}$$

Then by definition,

$$\begin{aligned}
\omega \wedge \eta &= \sum_{\hat{J},\hat{I}} \left(\sum_{\{I:\sigma_I(I)=\hat{I}\}} a_I(\boldsymbol{x}) \operatorname{sgn}(\sigma_I) \right) \cdot \\
&\qquad \left(\sum_{\{I:\sigma_J(J)=\hat{J}\}} b_J(\boldsymbol{x}) \operatorname{sgn}(\sigma_J) \right) d\boldsymbol{x}^{\hat{I}} \wedge d\boldsymbol{x}^{\hat{J}}
\end{aligned}$$

$$= \sum_{\hat{J},\hat{I}} \sum_{\{I:\sigma_I(I)=\hat{I}\}} \sum_{\{I:\sigma_J(J)=\hat{J}\}} a_I(\boldsymbol{x})\, b_J(\boldsymbol{x}) \operatorname{sgn}(\sigma_I) (\operatorname{sgn}(\sigma_J))\, d\boldsymbol{x}^{\hat{I}} \wedge d\boldsymbol{x}^{\hat{J}}$$

By the alternating property of determinants,

$$\begin{aligned}
&= \sum_{\hat{J},\hat{I}} \sum_{\{I:\sigma_I(I)=\hat{I}\}} \sum_{\{I:\sigma_J(J)=\hat{J}\}} a_I(\boldsymbol{x})\, b_J(\boldsymbol{x})\, d\boldsymbol{x}^I \wedge d\boldsymbol{x}^J \\
&= \sum_{I,J} a_I(\boldsymbol{x})\, b_J(\boldsymbol{x})\, d\boldsymbol{x}^I \wedge d\boldsymbol{x}^J \blacksquare
\end{aligned}$$

From this lemma, it is obvious that $\wedge$ acts like multiplication in the sense that it is distributive over addition, and associative. However, it is not commutative. From properties of determinants,

$$\begin{aligned}
& d\boldsymbol{x}^I \wedge d\boldsymbol{x}^J \\
= \;& dx_{i_1} \wedge \cdots \wedge dx_{i_k} \wedge dx_{j_1} \wedge \cdots \wedge dx_{j_l} \\
= \;& (-1)^{l+k-1}\, dx_{j_l} \wedge dx_{i_1} \wedge \cdots \wedge dx_{i_k} \wedge dx_{j_1} \wedge \cdots \wedge dx_{j_{l-1}} \\
= \;& (-1)^{l+k-1} (-1)^{l+k-1}\, dx_{j_{l-1}} \wedge dx_{j_l} \wedge dx_{i_1} \wedge \cdots \wedge dx_{i_k} \wedge dx_{j_1} \wedge \cdots \wedge dx_{j_{l-2}}
\end{aligned}$$

etc. Thus you get eventually

$$\left((-1)^{l+k-1} \right)^l dx_{j_1} \wedge \cdots \wedge dx_{j_l} \wedge dx_{i_1} \wedge \cdots \wedge dx_{i_k}$$

Now $l^2 - l + lk = lk + l(l-1)$ However, $l(l-1)$ must be even. Therefore,

$$= (-1)^{|I||J|} d\boldsymbol{x}^J \wedge d\boldsymbol{x}^I$$

where $|I|$ is the number of indices in I, similarly for J. This shows the following theorem which summarizes the algebraic properties of the wedge product.

Theorem 18.1.3 *Let α, β, γ be in some Ω^k. Then the following properties hold. If α, β are the same size,*

$$(\alpha + \beta) \wedge \gamma = \alpha \wedge \gamma + \beta \wedge \gamma$$

For α, β, γ arbitrary,

$$(\alpha \wedge \beta) \wedge \gamma = \alpha \wedge (\beta \wedge \gamma)$$

Finally, there is the condition about what happens when order is reversed. If $\alpha \in \Omega^k$ and $\beta \in \Omega^l$,

$$\alpha \wedge \beta = (-1)^{lk} \beta \wedge \alpha$$

Proof: The only claim which is not obvious is the last. However, this is also clear from Lemma 18.1.2 and the above computation involving $d\boldsymbol{x}^I \wedge d\boldsymbol{x}^J$. ■

18.2 The Exterior Derivative

A zero form is a function. Say f is such a smooth function. Then $df \in \Omega^1$ is defined as follows.

$$df \equiv \sum_{i=1}^{p} f_{x_i}(\boldsymbol{x})\, dx_i$$

Then for $\boldsymbol{r} \in C^1([0,1], \mathbb{R}^p)$,

$$\int_{\boldsymbol{r}} df \equiv \int_0^1 \sum_{i=1}^{p} f_{x_i}(\boldsymbol{r}(u))\, r_i'(u)\, du = \int_0^1 \frac{d}{du}(f(\boldsymbol{r}(u)))\, du = f(\boldsymbol{r}(1)) - f(\boldsymbol{r}(0))$$

In general, if you have something in Ω^k, you can define its exterior derivative as follows.

Definition 18.2.1 *Let $\omega = \sum_{I \text{ ascending}} a_I(\boldsymbol{x})\, d\boldsymbol{x}^I$. Then*

$$d\omega \equiv \sum_{I \text{ ascending}} da_I(\boldsymbol{x}) \wedge d\boldsymbol{x}^I$$

It is clear that d is linear.

It doesn't matter whether ω is written in terms of ascending indices. The same formula holds with no change.

Lemma 18.2.2 *Let $\omega = \sum_I a_I(\boldsymbol{x})\, d\boldsymbol{x}^I$ then $d\omega = \sum_I da_I(\boldsymbol{x}) \wedge d\boldsymbol{x}^I$*

Proof: Denote by $\hat{I}$ the ascending indices. Then if $\sigma_I(I)$ is ascending,

$$\omega = \sum_{\hat{I}} \sum_{\{I\} = \{\hat{I}\}} a_I(\boldsymbol{x})\, d\boldsymbol{x}^I = \sum_{\hat{I}} \left(\sum_{\{I\} = \{\hat{I}\}} a_I(\boldsymbol{x}) \operatorname{sgn}(\sigma_I(I)) \right) d\boldsymbol{x}^{\hat{I}}$$

Then since d is linear,

$$d\omega = \sum_{\hat{I}} \left(\sum_{\{I\}=\{\hat{I}\}} da_I(\boldsymbol{x}) \operatorname{sgn}(\sigma_I(I)) \right) \wedge d\boldsymbol{x}^{\hat{I}}$$

Then since $\wedge$ is also linear,

$$= \sum_{\hat{I}} \left(\sum_{\{I\}=\{\hat{I}\}} da_I(\boldsymbol{x}) \operatorname{sgn}(\sigma_I(I)) \wedge d\boldsymbol{x}^{\hat{I}} \right) = \sum_{\hat{I}} \left(\sum_{\{I\}=\{\hat{I}\}} da_I(\boldsymbol{x}) \wedge d\boldsymbol{x}^I \right)$$

$$= \sum_I da_I(\boldsymbol{x}) \wedge d\boldsymbol{x}^I \blacksquare$$

Next is a product rule. First note that it follows right away from the definition and the product rule from beginning calculus that

$$d(fg) = d(f)g + gd(g)$$

Lemma 18.2.3 *Let α, β be in Ω^k and Ω^l respectively. Then $d(\alpha \wedge \beta) = d\alpha \wedge \beta + (-1)^k \alpha \wedge d\beta$. Also $d^2 = 0$.*

Proof: Let $\alpha = \sum_I a_I(\boldsymbol{x}) d\boldsymbol{x}^I, \beta = \sum_J b_J(\boldsymbol{x}) d\boldsymbol{x}^J$. Then

$$\alpha \wedge \beta = \sum_{I,J} a_I b_J d\boldsymbol{x}^I \wedge d\boldsymbol{x}^J$$

and so $d(\alpha \wedge \beta)$ equals, thanks to the above lemma,

$$\sum_{I,J} d(a_I b_J) \wedge d\boldsymbol{x}^I \wedge d\boldsymbol{x}^J = \sum_{I,J} [d(a_I(\boldsymbol{x})) b_J(\boldsymbol{x}) + a_I(\boldsymbol{x}) d(b_J(\boldsymbol{x}))] \wedge d\boldsymbol{x}^I \wedge d\boldsymbol{x}^J$$

$$= \sum_{I,J} d(a_I(\boldsymbol{x})) b_J(\boldsymbol{x}) \wedge d\boldsymbol{x}^I \wedge d\boldsymbol{x}^J + \sum_{I,J} a_I(\boldsymbol{x}) d(b_J(\boldsymbol{x})) \wedge d\boldsymbol{x}^I \wedge d\boldsymbol{x}^J$$

From the definition of the wedge product and Lemma 18.1.2,

$$= d\alpha(\boldsymbol{x}) \wedge \beta + \sum_{I,J} a_I(\boldsymbol{x}) d(b_J(\boldsymbol{x})) \wedge d\boldsymbol{x}^I \wedge d\boldsymbol{x}^J$$

Now we will interchange the 1 form $d(b_J(\boldsymbol{x}))$ and the k form $d\boldsymbol{x}^I$. From Theorem 18.1.3

$$= d\alpha(\boldsymbol{x}) \wedge \beta(\boldsymbol{x}) + (-1)^k \sum_{I,J} a_I(\boldsymbol{x}) d\boldsymbol{x}^I \wedge d(b_J(\boldsymbol{x})) \wedge d\boldsymbol{x}^J$$

Then from Lemma 18.1.2 again,

$$\begin{aligned} &= d\alpha(\boldsymbol{x}) \wedge \beta(\boldsymbol{x}) + (-1)^k \sum_I a_I(\boldsymbol{x}) d\boldsymbol{x}^I \wedge \sum_J d(b_J(\boldsymbol{x})) \wedge d\boldsymbol{x}^J \\ &= d\alpha(\boldsymbol{x}) \wedge \beta(\boldsymbol{x}) + (-1)^k \alpha(\boldsymbol{x}) \wedge d\beta(\boldsymbol{x}) \end{aligned}$$

One of the important properties of the exterior derivative is that $d^2 = 0$. Let

$$\omega = \sum_I a_I(\boldsymbol{x}) d\boldsymbol{x}^I.$$

$$d\omega = \sum_I \sum_{r=1}^{p} a_{I,x_r} dx_r \wedge d\boldsymbol{x}^I$$

Then by definition,

$$\begin{aligned} d^2\omega &= \sum_I \sum_{r=1}^{p} \sum_{s=1}^{p} a_{I,x_r x_s} dx_s \wedge dx_r \wedge d\boldsymbol{x}^I \\ &= \sum_I \sum_{r<s} a_{I,x_r x_s} dx_s \wedge dx_r \wedge d\boldsymbol{x}^I + \sum_I \sum_{s<r} a_{I,x_r x_s} dx_s \wedge dx_r \wedge d\boldsymbol{x}^I \\ &= \sum_I \sum_{r<s} a_{I,x_r x_s} dx_s \wedge dx_r \wedge d\boldsymbol{x}^I + \sum_I \sum_{r<s} a_{I,x_s x_r} dx_r \wedge dx_s \wedge d\boldsymbol{x}^I \end{aligned}$$

In keeping with the convention that we assume a_I are as smooth as desired, we can conclude that the mixed partial derivatives are equal and so the above reduces to

$$\begin{aligned} & \sum_I \sum_{r<s} a_{I,x_r x_s} dx_s \wedge dx_r \wedge d\boldsymbol{x}^I + \sum_I \sum_{r<s} a_{I,x_r x_s} dx_r \wedge dx_s \wedge d\boldsymbol{x}^I \\ = & \sum_I \sum_{r<s} a_{I,x_r x_s} dx_s \wedge dx_r \wedge d\boldsymbol{x}^I - \sum_I \sum_{r<s} a_{I,x_r x_s} dx_s \wedge dx_r \wedge d\boldsymbol{x}^I = 0 \ \blacksquare \end{aligned}$$

It might be interesting to note that if one means weak derivatives, then the mixed partial derivatives are always equal.

18.3 Stokes Theorem

Now that the algebra of differential forms has been presented, it is time for the main topic Stokes theorem. Recall $[\boldsymbol{a},\boldsymbol{b}]$ is defined as $\equiv \prod_{l=1}^{k} [a_l, b_l]$. Let $\boldsymbol{x} \in \mathbb{R}^p, p \geq k, \boldsymbol{r} : [\boldsymbol{a},\boldsymbol{b}] \to \mathbb{R}^p$, and let ω a $k-1$ form be given as follows

$$\omega = \sum_{I \in J} \alpha_I(\boldsymbol{x})\, dx_{i_1} \wedge \cdots \wedge dx_{i_{k-1}}$$

where here $I = (i_1, ..., i_{k-1})$ is an increasing list of $k-1$ indices from $(1, 2, ..., p)$ and J will denote the set of all such increasing lists of indices. Thus there are $C(p, k-1)$ elements in the set J. Assume that α_I is C^1 but here $\boldsymbol{x} = \boldsymbol{r}(\boldsymbol{u})$ is C^2. After this case is done, it is easy to generalize.

This is a generalization of line integrals which is about integration over curves in space. Recall there was a parameter domain $[a,b]$ and a map $\boldsymbol{r} : [a,b] \to \mathbb{R}^p$ and there were two orientations or directions over the curve which was the set of points $\boldsymbol{r}([a,b])$. This concept of orientation is dealt with in multiple dimensions by the use of differential forms and the concept of determinants from linear algebra. The interval $[a,b]$ is replaced by $[\boldsymbol{a},\boldsymbol{b}]$. Stokes theorem then is a statement about $\boldsymbol{r}([\boldsymbol{a},\boldsymbol{b}])$ and the boundary of $\boldsymbol{r}([\boldsymbol{a},\boldsymbol{b}])$ just as it is in the case of a line integral.

Stoke's theorem actually follows from a computation.

$$d\omega \equiv \sum_{I \in J} \sum_{j=1}^{p} \frac{\partial \alpha_I}{\partial x_j}(\boldsymbol{x})\, dx_j \wedge dx_{i_1} \wedge \cdots \wedge dx_{i_{k-1}}$$

Definition 18.3.1 *As discussed earlier, we define*

$$\int_{\boldsymbol{r}} d\omega \equiv \sum_{I \in J} \sum_{j=1}^{p} \int_{[\boldsymbol{a},\boldsymbol{b}]} \frac{\partial \alpha_I}{\partial x_j}(\boldsymbol{r}(\boldsymbol{u})) \frac{\partial\left(x_j, x_{i_1} \cdots x_{i_{k-1}}\right)}{\partial(u_1, \cdots, u_k)} d\boldsymbol{u} \tag{18.3}$$

By definition, $\frac{\partial\left(x_j,x_{i_1}\cdots x_{i_{k-1}}\right)}{\partial(u_1,\cdots,u_k)}$ is the determinant of

$$\begin{pmatrix} x_{j,u_1} & x_{j,u_2} & \cdots & x_{j,u_k} \\ x_{i_1,u_1} & x_{i_1,u_2} & \cdots & x_{i_1,u_k} \\ \vdots & \vdots & & \vdots \\ x_{i_{k-1},u_1} & x_{i_{k-1},u_2} & \cdots & x_{i_{k-1},u_k} \end{pmatrix} \tag{18.4}$$

Note how this matrix is just the matrix of $D\boldsymbol{f}$ where

$$\boldsymbol{f}(\boldsymbol{u}) = \left(x_j(\boldsymbol{u}), x_{i_1}(\boldsymbol{u}), ..., x_{i_{k-1}}(\boldsymbol{u})\right)^T.$$

Then expanding the determinant in 18.3 along the first row, it equals

$$= \sum_{I\in J}\sum_{j=1}^{p}\int_{[\boldsymbol{a},\boldsymbol{b}]} \frac{\partial\alpha_I}{\partial x_j}(\boldsymbol{r}(\boldsymbol{u})) \overbrace{\sum_{l=1}^{k}\frac{\partial x_j}{\partial u_l}A^I_{1l}}^{\text{expanding determinant}} d\boldsymbol{u}$$

where A^I_{1l} is the $(1,l)^{th}$ cofactor for the determinant

$$A^I_{1l} = (-1)^{1+l}\frac{\partial\left(x_{i_1},\cdots,x_{i_{k-1}}\right)}{\partial\left(u_1,\cdots,\hat{u}_l\cdots,u_k\right)}, I = (i_1,\cdots,i_{k-1}) \tag{18.5}$$

Then this equals

$$= \sum_{I\in J}\sum_{l=1}^{k}\int_{[\boldsymbol{a},\boldsymbol{b}]}\sum_{j=1}^{p}\frac{\partial\alpha_I}{\partial x_j}(\boldsymbol{r}(\boldsymbol{u}))\frac{\partial x_j}{\partial u_l}A^I_{1l}d\boldsymbol{u} = \sum_{I\in J}\sum_{l=1}^{k}\int_{[\boldsymbol{a},\boldsymbol{b}]}\frac{\partial\alpha_I(\boldsymbol{r}(\boldsymbol{u}))}{\partial u_l}A^I_{1l}d\boldsymbol{u}$$

Now

$$\begin{aligned}\sum_l \frac{\partial\alpha_I(\boldsymbol{r}(\boldsymbol{u}))}{\partial u_l}A^I_{1l} &= \sum_l\frac{\partial}{\partial u_l}\left(\alpha_I(\boldsymbol{r}(\boldsymbol{u}))A^I_{1l}\right) - \sum_l \alpha_I(\boldsymbol{r}(\boldsymbol{u}))A^I_{1l,l} \\ &= \sum_l\frac{\partial}{\partial u_l}\left(\alpha_I(\boldsymbol{r}(\boldsymbol{u}))A^I_{1l}\right)\end{aligned}$$

By Lemma 7.11.2, that cofactor identity depending on equality of mixed partials. Therefore, from 18.3 and Fubini's theorem,

$$\begin{aligned}\int_{\boldsymbol{r}} d\omega &= \sum_{I\in J}\sum_{l=1}^{k}\int_{[\boldsymbol{a},\boldsymbol{b}]}\frac{\partial}{\partial u_l}\left(\alpha_I(\boldsymbol{r}(\boldsymbol{u}))A^I_{1l}\right) \\ &= \sum_{I\in J}\sum_{l=1}^{k}\int_{[\boldsymbol{a},\boldsymbol{b}]_l}\int_{[a_l,b_l]}\frac{\partial}{\partial u_l}\left(\left(\alpha_I(\boldsymbol{r}(\boldsymbol{u}))A^I_{1l}\right)\right)du_l d\boldsymbol{u}_l\end{aligned}$$

$$= \sum_{l=1}^{k}\int_{[\boldsymbol{a},\boldsymbol{b}]_l}\sum_{I\in J}\left(\left(\alpha_I\circ\boldsymbol{r}\right)A^I_{1l}\right)\left(\boldsymbol{u}_l(b_l)\right) - \left(\left(\alpha_I\circ\boldsymbol{r}\right)A^I_{1l}\right)\left(\boldsymbol{u}_l(a_l)\right)d\boldsymbol{u}_l \tag{18.6}$$

where here $[\boldsymbol{a},\boldsymbol{b}]_l$ means the $[a_l,b_l]$ is missing in the product $[\boldsymbol{a},\boldsymbol{b}]$ and

$$\boldsymbol{u}_l(b_l)$$

is $(u_1,...,u_{l-1},b_l,u_{l+1},...,u_k)$ with $\boldsymbol{u}_l(a_l)$ defined similarly. The term A^I_{1l} is the cofactor in 18.5 $(-1)^{1+l}\frac{\partial\left(x_{i_1},\cdots,x_{i_{k-1}}\right)}{\partial(u_1,\cdots,\hat{u}_l\cdots,u_k)}$. The term $\int_{[\boldsymbol{a},\boldsymbol{b}]_l}\left((\alpha_I\circ\boldsymbol{r})A^I_{1l}\right)(\boldsymbol{u}_l(b_l))\,d\boldsymbol{u}_l$ is an integration over the variables corresponding to a face of $[\boldsymbol{a},\boldsymbol{b}]$ and so it is a kind of boundary term.

Definition 18.3.2 *Let the $k-1$ form $\omega=\sum_{I\in J}\alpha_I(\boldsymbol{x})\,dx_{i_1}\wedge\cdots\wedge dx_{i_{k-1}}$ and let $\boldsymbol{r}:[\boldsymbol{a},\boldsymbol{b}]\subseteq\mathbb{R}^k\to\mathbb{R}^p$. The symbol $\int_{\partial\boldsymbol{r}}\omega$ is defined as 18.6.*

$$\sum_{l=1}^{k}\int_{[\boldsymbol{a},\boldsymbol{b}]_l}\sum_{I\in J}\left((\alpha_I\circ\boldsymbol{r})A^I_{1l}\right)(\boldsymbol{u}_l(b_l))-\left((\alpha_I\circ\boldsymbol{r})A^I_{1l}\right)(\boldsymbol{u}_l(a_l))\,d\boldsymbol{u}_l$$

where

$$(-1)^{1+l}\frac{\partial\left(x_{i_1},\cdots,x_{i_{k-1}}\right)}{\partial\left(u_1,\cdots,\hat{u}_l\cdots,u_k\right)}=A^I_{1l},\ I=(i_1,...,i_{k-1}),\ i_1<i_2<\cdots i_{k-1}$$

With this definition, here is Stoke's theorem.

Theorem 18.3.3 *Let $\omega=\sum_I\alpha_I(\boldsymbol{x})\,dx_{i_1}\wedge\cdots\wedge dx_{i_{k-1}}$ be an $k-1$ form. Let $\boldsymbol{r}:[\boldsymbol{a},\boldsymbol{b}]\to\mathbb{R}^p, p\geq k$ be in $C^2([\boldsymbol{a},\boldsymbol{b}];\mathbb{R}^p)$. Then $\int_{\partial\boldsymbol{r}}\omega=\int_{\boldsymbol{r}}d\omega$.*

Note that there is no assumption that $D\boldsymbol{r}$ has nonzero determinant. Everything above is valid under an assumption that $\boldsymbol{r}$ is only C^2. However, it may be necessary to assume more if it is desired to obtain a geometrically meaningful description of something in $\mathbb{R}^p$. Remember from calculus how you worked with smooth curves where the parametrization was one to one having nonvanishing derivative. There were reasons for this. If the derivative vanishes, this often yields a pointy place in the curve resulting from the given parametrization so it would not deserve to be called a smooth curve. It is the same here. Since $D\boldsymbol{r}$ is allowed to vanish, one can have $\boldsymbol{r}([\boldsymbol{a},\boldsymbol{b}])$ many different kinds of sets. However, if you insist that $D\boldsymbol{r}$ be invertible the points on the box would be preserved by an application of the implicit function theorem.

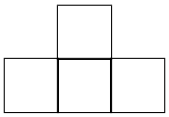
If you had $Q=[\boldsymbol{a},\boldsymbol{b}]$ and $\boldsymbol{y}=\boldsymbol{r}(\boldsymbol{x})$ for $\boldsymbol{x}\in U$, an open set containing Q and $\boldsymbol{y}\in\mathbb{R}^p, p>k$, then letting $\boldsymbol{F}(\boldsymbol{x},\boldsymbol{y})\equiv\boldsymbol{f}(\boldsymbol{x})-\boldsymbol{y}$, you could apply the implicit function theorem to $F(\boldsymbol{x},\boldsymbol{y})=\boldsymbol{0}$. If $D_1\boldsymbol{F}(\boldsymbol{x}_0,\boldsymbol{y}_0)$ has full rank n where $\boldsymbol{F}(\boldsymbol{x}_0,\boldsymbol{y}_0)=0$, could you smooth out the corners and edges of the box? Could you have $\boldsymbol{f}(S)$ where S is the part of the boundary of Q near $\boldsymbol{x}_0$ a corner of the box, part of a smooth surface? Certainly not because there would be smooth curves $u\to\boldsymbol{y}(u)$ lying in $\boldsymbol{f}(S)$ but then you could use the implicit function theorem and conclude that $\boldsymbol{x}=\boldsymbol{x}(\boldsymbol{y})$ is a C^1 function of $\boldsymbol{y}$ and so you would have $u\to\boldsymbol{x}(\boldsymbol{y}(u))$ a smooth curve in S which is not possible if S contains a corner. However, we do not assume that $D\boldsymbol{r}$ has full rank and so it would be possible for $\boldsymbol{r}(S)$ for S part of the boundary of Q to be smooth.

This could be improved by using the above to approximate functions which are not C^2 with functions which are, obtained by mollifying, and then passing to a limit to get more general situations. In particular, consider $\boldsymbol{r}$ a function in $C^1([\boldsymbol{a},\boldsymbol{b}];\mathbb{R}^p)$. Then by Lemma 16.4.1 there is a sequence of functions $\{\boldsymbol{r}_n\}$ each C^2 which converges uniformly to $\boldsymbol{r}$ and such that $D\boldsymbol{r}_n$ converges uniformly to $D\boldsymbol{r}$ on $[\boldsymbol{a},\boldsymbol{b}]$. Then Stoke's theorem holds with $\boldsymbol{r}$ replaced with $\boldsymbol{r}_n$ and so, passing to a limit as $n\to\infty$ one obtains Stoke's theorem for $\boldsymbol{r}$. This yields the following corollary.

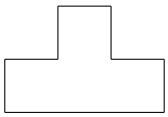
Corollary 18.3.4 *Let $\omega = \sum_I \alpha_I(\boldsymbol{x})\, dx_{i_1} \wedge \cdots \wedge dx_{i_{k-1}}$ be an $k-1$ form, each α_I being $C^1([\boldsymbol{a},\boldsymbol{b}])$. Let $\boldsymbol{r} : [\boldsymbol{a},\boldsymbol{b}] \to \mathbb{R}^p, p \geq k$ be in $C^1([\boldsymbol{a},\boldsymbol{b}];\mathbb{R}^p)$. Then $\int_{\partial \boldsymbol{r}} \omega = \int_{\boldsymbol{r}} d\omega$.*

Proof: From the above, $\int_{\partial \boldsymbol{r}_n} \omega = \int_{\boldsymbol{r}_n} d\omega$ where $\boldsymbol{r}_n$ is C^2. Both terms involve integrals over $[\boldsymbol{a},\boldsymbol{b}]$ or $[\boldsymbol{a},\boldsymbol{b}]_l$ and the convergence is uniform, so one can pass to a limit as $n \to \infty$ retaining the same formula with $\boldsymbol{r}_n$ replaced with $\boldsymbol{r}$. ■

Suppose you have two boxes $[\boldsymbol{a},\boldsymbol{b}]$ and $[\boldsymbol{c},\boldsymbol{d}]$ and these intersect on a common face, say the l^{th} face. Thus $c_l = b_l$. Then in the above description for the boundary integral on the common face, the two contributions cancel because you have the same thing except one has $|_{a_l}^{b_l}$ and the other has $|_{b_l}^{d_l}$. Therefore, you would get Stokes theorem for the box consisting of these two pasted together, the boundary integrals consisting of the sum of the boundary integrals of the remaining faces. Continuing this way, consider a chain of these boxes such that each box intersects another along a complete face. Then you could do the above for each pair of boxes in the chain and note that the boundary integrals will cancel along the common faces. Thus, in place of a single box you could have a much more complicated shape and the boundary integral would take place over exactly those faces which do not have intersection with faces of other boxes whereas $\int_{\boldsymbol{r}} d\omega$ would take place over the union of the boxes since the boundaries are sets of measure zero relative to m_k. The following picture illustrates what is meant.



Thus, adding over the boxes yields a parameter domain which looks like the following for $\int_{\boldsymbol{r}} d\omega$.



By now, it should be clear that fairly general regions can be included. Also, we only need to have a_I continuous on the face of two of these intersecting boxes. This is the analog of a piece-wise smooth curve.

18.4 Lipschitz Maps

This will be based on the approximation with C^1 maps. Let $\boldsymbol{r} : [\boldsymbol{a},\boldsymbol{b}] \subseteq \mathbb{R}^k \to \mathbb{R}^p$ where $\boldsymbol{r}$ is Lipschitz and $k \leq p$ as before. The idea is to extend the above Stokes theorem to this case where $\boldsymbol{r}$ is Lipschitz rather than C^1. First extend $\boldsymbol{r}$ as follows. For $t \geq b_l$ and $\boldsymbol{u}_l \in [\boldsymbol{a},\boldsymbol{b}]_l$, let

$$\boldsymbol{r}(u_1,...,u_{l-1},t,u_{l+1},...,u_k) \equiv \boldsymbol{r}(u_1,...,u_{l-1},b_l,u_{l+1},...,u_k)$$

and also if $t \leq a_l$ and $\boldsymbol{u}_l \in [\boldsymbol{a},\boldsymbol{b}]_l$, let

$$\boldsymbol{r}(u_1,...,u_{l-1},t,u_{l+1},...,u_k) \equiv \boldsymbol{r}(u_1,...,u_{l-1},a_l,u_{l+1},...,u_k)$$

This is done for each l. Then define for $h>0$

$$\boldsymbol{r}^h(\boldsymbol{u}) \equiv \left(\frac{1}{2h}\right)^k \int_{-2h+u_1+\frac{2h}{(b_1-a_1)}(u_1-a_1)}^{u_1+\frac{2h}{(b_1-a_1)}(u_1-a_1)} \cdots \int_{-2h+u_k+\frac{2h}{(b_k-a_k)}(u_k-a_k)}^{u_k+\frac{2h}{(b_k-a_k)}(u_k-a_k)} \boldsymbol{r}(\boldsymbol{t})\, dt_k \cdots dt_1 \quad (18.7)$$

Consider those integrals. When $u_1 = a_1$ you are integrating over $[a_1-2h, a_1]$ and when $u_1 = b_1$ you are integrating over $[b_1, b_1+2h]$. Of course it is similar for the other $[a_l, b_l]$. In general, each iterated integral is taken over an interval of length $2h$. For example,

$$u_1 + \frac{2h}{(b_1-a_1)}(u_1-a_1) - \left(-2h+u_1+\frac{2h}{(b_1-a_1)}(u_1-a_1)\right) = 2h$$

Now consider the right end of the l^{th} interval in $[\boldsymbol{a},\boldsymbol{b}]$ where $u_l = b_l$. Then this is describing one of the two faces corresponding to the l^{th} interval. By Fubini's theorem and the construction of the extention of $\boldsymbol{r}$, 18.7 implies that this equation simplifies to

$$\boldsymbol{r}^h(\boldsymbol{u}_l(b_l)) = \left(\frac{1}{2h}\right)^{k-1} \int_{-2h+u_1+\frac{2h}{(b_1-a_1)}(u_1-a_1)}^{u_1+\frac{2h}{(b_1-a_1)}(u_1-a_1)} \cdots \int_{-2h+u_k+\frac{2h}{(b_k-a_k)}(u_k-a_k)}^{u_k+\frac{2h}{(b_k-a_k)}(u_k-a_k)} \cdot$$

$$\boldsymbol{r}(t_1,...,b_l,...t_k)\, dt_k \cdots \widehat{dt_l} \cdots dt_1$$

Now by the fundamental theorem of calculus, for $i \neq l$,

$$\boldsymbol{r}^h_{u_i}(\boldsymbol{u}_l(b_l)) = \left(\frac{1}{2h}\right)^{k-2} \int_{-2h+u_1+\frac{2h}{(b_1-a_1)}(u_1-a_1)}^{u_1+\frac{2h}{(b_1-a_1)}(u_1-a_1)} \cdots \int_{-2h+u_k+\frac{2h}{(b_k-a_k)}(u_k-a_k)}^{u_k+\frac{2h}{(b_k-a_k)}(u_k-a_k)} \cdot$$

$$\boldsymbol{r}\left(t_1,...,u_i+\frac{2h}{(b_i-a_i)}(u_i-a_1),...b_l,...t_k\right)\frac{1}{2h}\left(1+\frac{2h}{b_i-a_i}\right) -$$

$$\boldsymbol{r}\left(t_1,...,-2h+u_i+\frac{2h}{(b_i-a_i)}(u_i-a_i),...b_l,...t_k\right)\frac{1}{2h}\left(1+\frac{2h}{b_i-a_i}\right)\cdot$$

$$dt_k \cdots \widehat{dt_i}, \cdots, \widehat{dt_l} \cdots dt_1$$

By the material on Rademacher's theorem, that integrand is

$$\left(1+\frac{2h}{b_i-a_i}\right)\frac{1}{2h}\int_{-2h+u_i+\frac{2h}{(b_i-a_i)}(u_i-a_i)}^{u_i+\frac{2h}{(b_i-a_i)}(u_i-a_i)} \boldsymbol{r}_{u_i}(t_1,...,t_i,...b_l,...t_k)\, dt_i$$

Now a.e. point of $[\boldsymbol{a},\boldsymbol{b}]_l$ is a Lebesgue point and so we can pass to a limit and obtain pointwise a.e. convergence of $\boldsymbol{r}^h_{u_i}(\boldsymbol{u}_l(b_l))$ to $\boldsymbol{r}_{u_i}(\boldsymbol{u}_l(b_l))$.

Some comment on this might be useful because these Lebesgue points are not always at the center of the box of sides of length $2h$ determined by the limits of the iterated integrals. A given point $\boldsymbol{u}_l(b_l) \in B(\boldsymbol{u}_l(b_l), 3h)$ where this ball is taken with respect to $\|\cdot\|_\infty$. Thus, from Lebesgue's fundamental theorem of calculus, we have on this face at Lebesgue points the following converges to 0 as $h \to 0$ which is what was desired

$$\left(1+\frac{2h}{b_i-a_i}\right)\left(\frac{3}{2}\right)^{k-1}\cdot$$

$$\left(\frac{1}{3h}\right)^{k-1}\int_{B(\boldsymbol{u}_l(b_l),3h)} \|\boldsymbol{r}_{u_i}(\boldsymbol{u}_l(b_l)) - \boldsymbol{r}_{u_i}(t_1,...,t_i,...b_l,...t_k)\|\, d\boldsymbol{t}_l$$

Indeed, $\left(\frac{3}{2}\right)^{k-1}\left(\frac{1}{3h}\right)^{k-1}=\left(\frac{1}{2h}\right)^{k-1}$ and the integral in the above is taken over the larger set $B(\boldsymbol{u}_l(b_l),3h)$.

As to points of $[\boldsymbol{a},\boldsymbol{b}]$, the pointwise convergence of $\boldsymbol{r}^h_{u_j}(\boldsymbol{u})$ to $\boldsymbol{r}_{u_j}(\boldsymbol{u})$ follows from similar reasoning but is a little less involved. Now $\boldsymbol{r}^h$ is clearly C^1 and so we have Stokes theorem for $\boldsymbol{r}^h$.

$$\begin{aligned}&\sum_{l=1}^{k}\int_{[\boldsymbol{a},\boldsymbol{b}]_l}\sum_{I\in J}\left(\left(\alpha_I\circ\boldsymbol{r}^h\right)A^{Ih}_{1l}\right)(\boldsymbol{u}_l(b_l))-\left((\alpha_I\circ\boldsymbol{r})A^{Ih}_{1l}\right)(\boldsymbol{u}_l(a_l))\,d\boldsymbol{u}_l\\ =\;&\int_{\boldsymbol{r}}d\omega\equiv\sum_{I\in J}\sum_{j=1}^{p}\int_{[\boldsymbol{a},\boldsymbol{b}]}\frac{\partial\alpha_I}{\partial x_j}\left(\boldsymbol{r}^h(\boldsymbol{u})\right)\frac{\partial\left(x_j,x_{i_1}\cdots x_{i_{k-1}}\right)}{\partial(u_1,\cdots,u_k)}d\boldsymbol{u}\end{aligned}$$

where the superscript indicates that all is defined in terms of $\boldsymbol{r}^h$. Then from the dominated convergence theorem, it follows that we can pass to a limit and obtain Stokes theorem where the boundary terms are defined from Rademacher's theorem on the faces of $[\boldsymbol{a},\boldsymbol{b}]$.

Theorem 18.4.1 *Let*

$$\boldsymbol{r}:[\boldsymbol{a},\boldsymbol{b}]\subseteq\mathbb{R}^k\to\mathbb{R}^p,p\geq k$$

, be Lipschitz and $\alpha_I\in C^1(\boldsymbol{r}([\boldsymbol{a},\boldsymbol{b}]))$ for $I\subseteq J$, the set of increasing lists of $k-1$ indices from $(1,2,...,p)$. Then one obtains Stokes theorem

$$\begin{aligned}&\sum_{l=1}^{k}\int_{[\boldsymbol{a},\boldsymbol{b}]_l}\sum_{I\in J}\left((\alpha_I\circ\boldsymbol{r})A^{I}_{1l}\right)(\boldsymbol{u}_l(b_l))-\left((\alpha_I\circ\boldsymbol{r})A^{I}_{1l}\right)(\boldsymbol{u}_l(a_l))\,d\boldsymbol{u}_l\\ =\;&\int_{\boldsymbol{r}}d\omega\equiv\sum_{I\in J}\sum_{j=1}^{p}\int_{[\boldsymbol{a},\boldsymbol{b}]}\frac{\partial\alpha_I}{\partial x_j}(\boldsymbol{r}(\boldsymbol{u}))\frac{\partial\left(x_j,x_{i_1}\cdots x_{i_{k-1}}\right)}{\partial(u_1,\cdots,u_k)}d\boldsymbol{u}\end{aligned}$$

where the partial derivatives of $A^I_{1l}=(-1)^{1+l}\frac{\partial\left(x_{i_1},\cdots,x_{i_{k-1}}\right)}{\partial(u_1,\cdots,\hat{u}_l\cdots,u_k)}$ are defined in terms of Rademacher's theorem applied to the $k-1$ dimensional faces of $[\boldsymbol{a},\boldsymbol{b}]$.

18.5 What Does it Mean?

Stokes theorem is a statement about integration by parts. However, one can give geometric meaning to what it says. These considerations will come from the area formula which I will use as needed.

For a particular l there are two faces in the boundary term for the Stokes formula. Consider the one where the l^{th} component is b_l. Recall that J was the set of increasing lists of $k-1$ indices.

$$\int_{[\boldsymbol{a},\boldsymbol{b}]_l}\sum_{I\in J}\left((\alpha_I\circ\boldsymbol{r})A^{I}_{1l}\right)(\boldsymbol{u}_l(b_l))\,d\boldsymbol{u}_l$$

Here $A^I_{1l}=(-1)^{1+l}\frac{\partial\left(x_{i_1},\cdots,x_{i_{k-1}}\right)}{\partial(u_1,\cdots,\hat{u}_l\cdots,u_k)}$ and $I=(i_1,...,i_{k-1})$. Letting

$$J_*(\boldsymbol{u}_l)=\sqrt{\sum_{I\in J}\left(A^I_{1l}\right)^2(\boldsymbol{u}_l(b_l))},$$

this term is of the form

$$\int_{[\boldsymbol{a},\boldsymbol{b}]_l} \sum_{I\in J} \left((\alpha_I \circ \boldsymbol{r}) \frac{A^I_{1l}}{J_*(\boldsymbol{u}_l)} \right) (\boldsymbol{u}_l(b_l)) J_*(\boldsymbol{u}_l)\, d\boldsymbol{u}_l \tag{18.8}$$

Define $\frac{A^I_{1l}}{J_*(\boldsymbol{u}_l)} = 0$ if $J_*(\boldsymbol{u}_l) = 0$. By Lemma 17.4.1 Lipschitz $\boldsymbol{r}_{b_l}$ the restriction of $\boldsymbol{r}$ to the face of $[\boldsymbol{a},\boldsymbol{b}]$ corresponding to the l^{th} component equal to b_l, this mapping $\boldsymbol{r}_{b_l}$ maps this set to one which has $\mathscr{H}^{k-1}$ measure zero. Using the Binet Cauchy theorem to identify $J_*(\boldsymbol{u}_l)$ with $\left(\det\left(D\boldsymbol{r}_{b_l}(\boldsymbol{u}_l)^* D\boldsymbol{r}_{b_l}(\boldsymbol{u}_l)\right)\right)^{1/2}$, 18.8 reduces to

$$\int_{\boldsymbol{r}_{b_l}([\boldsymbol{a},\boldsymbol{b}]_l)} \#(\boldsymbol{x}) \sum_{I\in J} \alpha_I(\boldsymbol{x}) N^I_{b_l}(\boldsymbol{x})\, d\mathscr{H}^{k-1}(\boldsymbol{x})$$

where $N^I_{b_l}\left(\boldsymbol{r}_{b_l}(\boldsymbol{u}_l)\right) = \frac{A^I_{1l}}{J_*(\boldsymbol{u}_l)}(\boldsymbol{u}_l(b_l)) = \frac{1}{J_*(\boldsymbol{u}_l)}(-1)^{1+l} \frac{\partial\left(x_{i_1},\cdots,x_{i_{k-1}}\right)}{\partial(u_1,\cdots,\hat{u}_l\cdots,u_k)}$ is a component of a unit vector in $\mathbb{R}^{C(p,k-1)}$ at least $\mathscr{H}^{k-1}$ a.e. Assume that $\boldsymbol{r}_{b_l}$ is one to one or is one to one off a set S which has $\mathscr{H}^{k-1}\left(\boldsymbol{r}_{b_l}(S)\right) = 0$. That way we can eliminate $\#(\boldsymbol{x})$ the number of times $\boldsymbol{x}$ is hit by $\boldsymbol{r}_{b_l}$. Thus, generalizing the notation, the boundary term in Stokes theorem is of the form

$$\begin{aligned}
&\sum_{l=1}^k \int_{\boldsymbol{r}_{b_l}([\boldsymbol{a},\boldsymbol{b}]_l)} \sum_{I\in J} \alpha_I(\boldsymbol{x}) N^I_{b_l}(\boldsymbol{x})\, d\mathscr{H}^{k-1}(\boldsymbol{x}) \\
&-\sum_{l=1}^k \int_{\boldsymbol{r}_{a_l}([\boldsymbol{a},\boldsymbol{b}]_l)} \sum_{I\in J} \alpha_I(\boldsymbol{x}) N^I_{a_l}(\boldsymbol{x})\, d\mathscr{H}^{k-1}(\boldsymbol{x})
\end{aligned}$$

where $\sum_{I\in J}\left(N^I_{b_l}(\boldsymbol{x})\right)^2 = 1$. Letting $N^I = N^I_{b_l}$ on $\boldsymbol{r}_{b_l}([\boldsymbol{a},\boldsymbol{b}]_l)$ and $-N^I_{a_l}$ on $\boldsymbol{r}_{a_l}([\boldsymbol{a},\boldsymbol{b}]_l)$, the above is of the form

$$\int_{\boldsymbol{r}(\partial[\boldsymbol{a},\boldsymbol{b}])} \sum_{I\in J} \alpha_I(\boldsymbol{x}) N^I(\boldsymbol{x})\, d\mathscr{H}^{k-1}$$

Similarly

$$\begin{aligned}
\int_{\boldsymbol{r}} d\omega &\equiv \sum_{I\in J} \int_{[\boldsymbol{a},\boldsymbol{b}]} \sum_{j=1}^p \frac{\partial \alpha_I}{\partial x_j}(\boldsymbol{r}(\boldsymbol{u})) \frac{\partial\left(x_j, x_{i_1}\cdots x_{i_{k-1}}\right)}{\partial(u_1,\cdots,u_k)} d\boldsymbol{u} \\
&= \int_{[\boldsymbol{a},\boldsymbol{b}]} \sum_{j=1}^p \sum_{I\in J} \frac{\partial \alpha_I}{\partial x_j}(\boldsymbol{r}(\boldsymbol{u})) \frac{\partial\left(x_j, x_{i_1}\cdots x_{i_{k-1}}\right)}{\partial(u_1,\cdots,u_k)} d\boldsymbol{u}
\end{aligned}$$

Now that determinant is only nonzero if j is none of the i_s. By the Binet Cauchy theorem,

$$\det\left(D\boldsymbol{r}(\boldsymbol{u})^* D\boldsymbol{r}(\boldsymbol{u})\right) = J_*(\boldsymbol{u})^2 = \sum_{x_{i_1}<\cdots<x_{i_k}} \left(\frac{\partial\left(x_{i_1},\cdots,x_{i_k}\right)}{\partial(u_1,\cdots,u_k)} \right)^2$$

and so it follows from the area formula that there exists N^I_j for each I an increasing list of

$k-1$ indices such that for J all such increasing lists, $\sum_{j=1}^{p}\sum_{I\in J}\left(N_j^I\right)^2 = 1$ and

$$\begin{aligned}\int_{\boldsymbol{r}} d\omega &= \int_{[\boldsymbol{a},\boldsymbol{b}]} \sum_{j=1}^{p}\sum_{I\in J} \frac{\partial \alpha_I}{\partial x_j}(\boldsymbol{r}(\boldsymbol{u})) \frac{\frac{\partial\left(x_j,x_{i_1}\cdots x_{i_{k-1}}\right)}{\partial(u_1,\cdots,u_k)}}{J_*(\boldsymbol{u})} J_*(\boldsymbol{u})\,d\boldsymbol{u} \\ &= \int_{\boldsymbol{r}([\boldsymbol{a},\boldsymbol{b}])} \#(\boldsymbol{x}) \sum_{I\in J}\sum_{j=1}^{p} \frac{\partial \alpha_I}{\partial x_j}(\boldsymbol{x}) N_j^I(\boldsymbol{x})\, d\mathscr{H}^k \end{aligned} \tag{18.9}$$

where $\#(\boldsymbol{x})$ is the number of times $\boldsymbol{x}$ is hit by $\boldsymbol{r}$. Thus if $\boldsymbol{r}$ is one to one off $\hat{S}$ where $\boldsymbol{r}\left(\hat{S}\right)$ has $\mathscr{H}^k$ measure zero, it follows that we can remove $\#(\boldsymbol{x})$ from the formula. As before, we can ignore the set where $J_*(\boldsymbol{u}) = 0$ thanks to Lemma 17.4.1.

Observation 18.5.1 *Stokes theorem may be thought of as a statement about $\boldsymbol{r}([\boldsymbol{a},\boldsymbol{b}])$ and $\boldsymbol{r}(\partial[\boldsymbol{a},\boldsymbol{b}])$ which involves geometrical concepts dependent on these sets. This holds whenever Lipschitz $\boldsymbol{r}$ restricted to each face of $[\boldsymbol{a},\boldsymbol{b}]$ is one to one off a set S where $\boldsymbol{r}(S)$ has $\mathscr{H}^{k-1}$ measure zero, and $\boldsymbol{r}$ is one to one off $\hat{S}$ where $\boldsymbol{r}\left(\hat{S}\right)$ has $\mathscr{H}^k$ measure zero.*

I think that the most important case is where $k = p$ and in this case we have the divergence theorem. Here there are exactly $p-1$ increasing lists of indices I and these are of the form $\left(1,\cdots,\widehat{j},\cdots,p\right)$. We let α_I be denoted as $a_j(-1)^{j+1}$ where j is the index missing in I. Therefore, the formula for $\int_{\boldsymbol{r}} d\omega$ reduces to

$$\pm\int_{\boldsymbol{r}([\boldsymbol{a},\boldsymbol{b}])} \sum_{j=1}^{p} \frac{\partial a_j}{\partial x_j}(\boldsymbol{x})\, d\mathscr{H}^k$$

assuming that $\boldsymbol{r}$ is one to one off a set S for which $\boldsymbol{r}(S)$ has $\mathscr{H}^p$ measure zero. This is because from 18.9

$$\frac{\partial \alpha_I}{\partial x_j}(\boldsymbol{r}(\boldsymbol{u})) \frac{\frac{\partial\left(x_j,x_{i_1}\cdots x_{i_{k-1}}\right)}{\partial(u_1,\cdots,u_k)}}{J_*(\boldsymbol{u})} J_*(\boldsymbol{u}) = \left((-1)^{j+1}\right)^2 a_i(\boldsymbol{r}(\boldsymbol{u})) \frac{\frac{\partial\left(x_1,\cdots,x_j\cdots,x_p\right)}{\partial(u_1,\cdots,u_k)}}{J_*(\boldsymbol{u})} = \pm 1$$

The boundary terms reduce to

$$\begin{aligned}&\sum_{l=1}^{k}\int_{\boldsymbol{r}_{b_l}([\boldsymbol{a},\boldsymbol{b}]_l)} \sum_{j=1}^{p} a_j(\boldsymbol{x}) N^j(\boldsymbol{x})\, d\mathscr{H}^{k-1}(\boldsymbol{x}) \\ &-\sum_{l=1}^{k}\int_{\boldsymbol{r}_{a_l}([\boldsymbol{a},\boldsymbol{b}]_l)} \sum_{j=1}^{p} a_j(\boldsymbol{x}) N^j(\boldsymbol{x})\, d\mathscr{H}^{k-1}(\boldsymbol{x})\end{aligned}$$

where for $\boldsymbol{x} = \boldsymbol{r}(\boldsymbol{u}), N^j(\boldsymbol{x}) = (-1)^{j+l}\frac{\partial\left(x_1,\cdots,\hat{x}_j\cdots,x_p\right)}{\partial\left(u_1,\cdots,\hat{u}_l,\cdots,u_p\right)}$. This can be written in the form

$$\pm\int_{\boldsymbol{r}([\boldsymbol{a},\boldsymbol{b}])} \sum_{j=1}^{p} \frac{\partial a_j}{\partial x_j}(\boldsymbol{x})\, d\mathscr{H}^k = \int_{\boldsymbol{r}(\partial[\boldsymbol{a},\boldsymbol{b}])} \sum_{j=1}^{p} a_j(\boldsymbol{x}) N^j(\boldsymbol{x})\, d\mathscr{H}^{k-1}(\boldsymbol{x})$$

This divergence theorem is discussed more later when also the unit vector $\boldsymbol{N}$ whose j^{th} component is N^j is described as an exterior unit normal provided $\frac{\partial\left(x_1,\cdots,x_p\right)}{\partial(u_1,\cdots,u_k)} \geq 0$. Note how the $(-1)^{j+1}$ on the a_j is responsible for the $(-1)^{j+l}$ instead of $(-1)^{1+l}$ in the description of N^j.

18.6 Examples of $\boldsymbol{r}([\boldsymbol{a},\boldsymbol{b}])$

I want to point out that there are many examples of $\boldsymbol{r}([\boldsymbol{a},\boldsymbol{b}])$ which fit into the above integration by parts idea of Stokes theorem in which $\boldsymbol{r}$ is Lipschitz, in order to tie this more to the way we usually think of these theorems in Calculus. Consider the following picture in which a closed ball B of radius 1 is inscribed into the box $Q \equiv [-\mathbf{1},\mathbf{1}] \equiv \prod_{i=1}^{k} [-1,1]$. Let P be the projection map onto this ball B.

Then it is geometrically obvious that the projection map P satisfies $P(Q \setminus B) = \partial B$ a set of $\mathscr{H}^k$ measure zero and that $P : \partial[-\mathbf{1},\mathbf{1}] \to \partial B$ is one to one and onto on $\partial[-\mathbf{1},\mathbf{1}]$. Now let $\boldsymbol{r} : B \to \mathbb{R}^p$ for $p \geq k$ be Lipschitz and one to one. Then $\boldsymbol{r} \circ P : [-\mathbf{1},\mathbf{1}] \to \boldsymbol{r}(B)$ satisfies all the necessary conditions for an application of Stokes theorem including the geometric descriptions just given. What kinds of sets are in $\boldsymbol{r}(B)$ for B a closed ball and $\boldsymbol{r}$ Lipschitz and one to one? I think you can see that this would include virtually everything of interest. You could stretch B in various directions, pinch it, bend it, etc. Roughly speaking, imagine a ball of soft clay and doing what a child would do to it before he tears it into little pieces, throws them around the room and stomps them into the carpet. The result would be one of the possible sets $\boldsymbol{r}(B)$. Since $\boldsymbol{r}$ is a homeomorphism, the interior of B corresponds to the relative interior of $\boldsymbol{r}(B)$, points $\boldsymbol{x} \in \boldsymbol{r}(B) = \boldsymbol{r} \circ P([-\mathbf{1},\mathbf{1}])$ which are not in $\boldsymbol{r}(\partial B)$. The boundary faces of $[-\mathbf{1},\mathbf{1}]$ and $\boldsymbol{r} \circ P$ restricted to these faces will parametrize finitely many disjoint pieces of $\boldsymbol{r}(\partial B)$.

Of course you could also consider chains of such boxes as described earlier in the case that $\boldsymbol{r}$ is C^1. However, when you can allow $\boldsymbol{r}$ to be Lipschitz, it is clear that the theory is sufficiently general to include most things which would be of interest in any application from a single box. Next is a discussion of orientation placed here to make an analogy with the case of line integrals and oriented curves.

18.7 Orientation and Degree

Here I will consider orientation briefly. As in the case of a curve, it reduces to considerations of $\boldsymbol{r}^{-1} \circ \hat{\boldsymbol{r}}$.

Proposition 18.7.1 *Suppose* $\boldsymbol{r}([\boldsymbol{a},\boldsymbol{b}]) = \hat{\boldsymbol{r}}\left(\left[\hat{\boldsymbol{a}},\hat{\boldsymbol{b}}\right]\right)$*, two sets in* $\mathbb{R}^p$ *and both* $\boldsymbol{r},\hat{\boldsymbol{r}}$ *are one to one and Lipschitz,* $[\boldsymbol{a},\boldsymbol{b}], \left[\hat{\boldsymbol{a}},\hat{\boldsymbol{b}}\right]$ *being two parameter domains in* $\mathbb{R}^k, k \leq p$*. Then for*

$$\boldsymbol{\omega} = a(\boldsymbol{x})\, dx_{i_1} \wedge \cdots \wedge dx_{i_k}$$

Assume also that $\boldsymbol{r}^{-1} \circ \hat{\boldsymbol{r}}$ *is Lipschitz and* $\det\left(D\left(\boldsymbol{r}^{-1} \circ \hat{\boldsymbol{r}}\right)(\boldsymbol{t})\right) \geq 0$ *a.e. This is a statement about orientation. It follows then that* $\int_{\boldsymbol{r}} \boldsymbol{\omega} = \int_{\hat{\boldsymbol{r}}} \boldsymbol{\omega}$*.*

Proof: Let $(\boldsymbol{a},\boldsymbol{b}) \equiv \prod_{j=1}^{k} (a_j, b_j)$. Now from the area formula

$$\int_{\boldsymbol{r}} \boldsymbol{\omega} \equiv \int_{[\boldsymbol{a},\boldsymbol{b}]} a(\boldsymbol{r}(\boldsymbol{u})) \frac{\partial(x_{i_1},...,x_{i_k})}{\partial(u_1,...,u_k)}(\boldsymbol{u})\, du$$

$$= \int_{\left[\hat{a},\hat{b}\right]} a\left(\boldsymbol{r}\left(\boldsymbol{r}^{-1}\circ\hat{\boldsymbol{r}}\right)(\boldsymbol{t})\right)\frac{\partial\left(x_{i_1},\ldots,x_{i_k}\right)}{\partial\left(u_1,\ldots,u_k\right)}\left(\boldsymbol{r}^{-1}\circ\hat{\boldsymbol{r}}(\boldsymbol{t})\right)\det\left(D\left(\boldsymbol{r}^{-1}\circ\hat{\boldsymbol{r}}\right)(\boldsymbol{t})\right)dt$$

$$= \int_{\left[\hat{a},\hat{b}\right]} a\left(\boldsymbol{r}\left(\boldsymbol{r}^{-1}\circ\hat{\boldsymbol{r}}\right)(\boldsymbol{t})\right)\frac{\partial\left(x_{i_1},\ldots,x_{i_k}\right)}{\partial\left(t_1,\ldots,t_k\right)}(\boldsymbol{t})\,dt \equiv \int_{\hat{\boldsymbol{r}}}\boldsymbol{\omega}\ \blacksquare$$

An application of the area formula gives the following corollary. I will use $\boldsymbol{r}^{-1}\circ\hat{\boldsymbol{r}}$ to denote a Lipschitz function which is one to one off a set S which equals the Lipschitz function $\boldsymbol{r}^{-1}\circ\hat{\boldsymbol{r}}$ on S^C. In particular if S has measure 0 and $\boldsymbol{r}^{-1}\circ\hat{\boldsymbol{r}}$ is Lipschitz on S^C, then you could extend to a Lipschitz function which would map S to a set of measure zero, thus being in the situation of this corollary.

Corollary 18.7.2 *Suppose $\boldsymbol{r}^{-1}\circ\hat{\boldsymbol{r}}$ is one to one off a set $S\subseteq\left[\hat{\boldsymbol{a}},\hat{\boldsymbol{b}}\right]$ and that $\boldsymbol{r}^{-1}\circ\hat{\boldsymbol{r}}(S)$ has measure zero. Then the above would hold with no change.*

Since this allows for Lipschitz functions, this is slightly more general than the usual situation from Calculus even in one dimension. However, more can be said. Orientation is really a statement about the degree of the map $\boldsymbol{r}^{-1}\circ\hat{\boldsymbol{r}}$, a concept which makes perfect sense without any direct reference to differentiability.

Recall that with a smooth curve C having points $\boldsymbol{p},\boldsymbol{q}$ and a one to one map to this curve, there are two ways to move over the curve, from $\boldsymbol{p}$ to $\boldsymbol{q}$ or from $\boldsymbol{q}$ to $\boldsymbol{p}$. One defines equivalence classes on the continuous mappings $\boldsymbol{r}$ which map a closed interval to C. Two of these $\boldsymbol{r},\hat{\boldsymbol{r}}$ are equivalent if $\boldsymbol{r}^{-1}\circ\hat{\boldsymbol{r}}$ is increasing. It follows from the intermediate value theorem of Bolzano and a simple argument that this composition of maps is either increasing or decreasing. Thus, from the theorem about differentiation of monotone functions, $\left(\boldsymbol{r}^{-1}\circ\hat{\boldsymbol{r}}\right)'$ is nonnegative a.e. exactly when the two parametrizations give the same orientation. In the above, this is determined by $\det\left(D\left(\boldsymbol{r}^{-1}\circ\hat{\boldsymbol{r}}\right)\right)$. In addition, this reduces to a topological notion having to do with the degree. Instead of "increasing" we say that $d\left(\boldsymbol{r}^{-1}\circ\hat{\boldsymbol{r}},\Omega,(\boldsymbol{a},\boldsymbol{b})\right)$ is 1. The notion of "increasng" is not available.

Recall that from Proposition 15.6.7, Corollary 15.6.6, $d\left(\boldsymbol{r}^{-1}\circ\hat{\boldsymbol{r}},\left(\hat{\boldsymbol{a}},\hat{\boldsymbol{b}}\right),(\boldsymbol{a},\boldsymbol{b})\right)$ is either 1 or -1. This is the topological degree of the mapping $\boldsymbol{r}^{-1}\circ\hat{\boldsymbol{r}}$ which is constant on the connected component $(\boldsymbol{a},\boldsymbol{b})$ of $\mathbb{R}^k\setminus\boldsymbol{r}^{-1}\circ\hat{\boldsymbol{r}}\left(\partial\left(\left[\hat{\boldsymbol{a}},\hat{\boldsymbol{b}}\right]\right)\right)$. Suppose then that this degree $d\left(\boldsymbol{r}^{-1}\circ\hat{\boldsymbol{r}},\left(\hat{\boldsymbol{a}},\hat{\boldsymbol{b}}\right),\boldsymbol{y}\right)$ is 1. Then from Lemma 17.10.5 on Page 487, it follows that whenever $f\in C_c(\boldsymbol{a},\boldsymbol{b})$

$$\int f(\boldsymbol{u})\,d\left(\boldsymbol{r}^{-1}\circ\hat{\boldsymbol{r}},\Omega,\boldsymbol{u}\right)du = \int_{\left(\hat{\boldsymbol{a}},\hat{\boldsymbol{b}}\right)} f\left(\boldsymbol{r}^{-1}\circ\hat{\boldsymbol{r}}(\boldsymbol{t})\right)\det\left(D\left(\boldsymbol{r}^{-1}\circ\hat{\boldsymbol{r}}\right)(\boldsymbol{t})\right)dt$$

You could take an increasing sequence $f_n(\boldsymbol{u})\to\mathcal{X}_{(\boldsymbol{a},\boldsymbol{b})}(\boldsymbol{u})$ of the above sort. From the area formula,

$$\int f_n(\boldsymbol{u})\,du = \int_{\left(\hat{\boldsymbol{a}},\hat{\boldsymbol{b}}\right)} f\left(\boldsymbol{r}^{-1}\circ\hat{\boldsymbol{r}}(\boldsymbol{t})\right)\left|\det\left(D\left(\boldsymbol{r}^{-1}\circ\hat{\boldsymbol{r}}\right)(\boldsymbol{t})\right)\right|dt$$

and so

$$0 = \int_{\left(\hat{\boldsymbol{a}},\hat{\boldsymbol{b}}\right)} f_n\left(\boldsymbol{r}^{-1}\circ\hat{\boldsymbol{r}}(\boldsymbol{t})\right)\left(\left|\det\left(D\left(\boldsymbol{r}^{-1}\circ\hat{\boldsymbol{r}}\right)(\boldsymbol{t})\right)\right|-\det\left(D\left(\boldsymbol{r}^{-1}\circ\hat{\boldsymbol{r}}\right)(\boldsymbol{t})\right)\right)dt.$$

Using the monotone convergence theorem, it follows that

$$0=\int_{(\hat{\boldsymbol{a}},\hat{\boldsymbol{b}})}\left(\left|\det\left(D\left(\boldsymbol{r}^{-1}\circ\hat{\boldsymbol{r}}\right)(\boldsymbol{t})\right)\right|-\det\left(D\left(\boldsymbol{r}^{-1}\circ\hat{\boldsymbol{r}}\right)(\boldsymbol{t})\right)\right)dt.$$

and so $\det\left(D\left(\boldsymbol{r}^{-1}\circ\hat{\boldsymbol{r}}\right)(\boldsymbol{t})\right)\geq 0$ a.e. This is the condition given in the above proposition. Thus this condition which applies to Lipschitz functions follows from a statement that $d\left(\boldsymbol{r}^{-1}\circ\hat{\boldsymbol{r}},\Omega,\boldsymbol{u}\right)=1$. Conversely, if the condition $\det\left(D\left(\boldsymbol{r}^{-1}\circ\hat{\boldsymbol{r}}\right)(\boldsymbol{t})\right)\geq 0$ a.e. it will follow from the above formula that $d\left(\boldsymbol{r}^{-1}\circ\hat{\boldsymbol{r}},\Omega,\boldsymbol{u}\right)=1$ since it cannot equal -1.

18.8 Examples of Stoke's Theorem

Here the attempt is made to tie this formalism to the usual things studied in calculus.

18.8.1 Fundamental Theorem of Calculus

First let $k=p=1$. What does Stoke's theorem say? In this case, $d\omega$ is a 1 form and so ω should be a 0 form which is just a function.

$$\omega=a(x),\ d\omega=a'(x)\,dx$$

Then if $r:[a,b]\to\mathbb{R}$ is C^1,

$$\int_{\boldsymbol{r}}d\omega\equiv\int_{\boldsymbol{r}}a'(x)\,dx\equiv\int_a^b a'(r(u))\,\frac{dr}{du}du=a(r(b))-a(r(a))$$

$$\int_{\partial\boldsymbol{r}}\omega=\int_{\partial\boldsymbol{r}}a(x)=(-1)^2a(r(b))-(-1)^2a(r(a))$$

which is the same thing. It is just the fundamental theorem of calculus essentially.

18.8.2 Line Integrals for Conservative Fields

What if $p=3,k=1$? This is similar. You need to have ω a zero form. Thus $\omega=a(x_1,...,x_m)$. Then $d\omega=\sum_i a_{x_i}dx_i$. Letting $\boldsymbol{r}:[a,b]\to\mathbb{R}^m$,

$$\int_{\boldsymbol{r}}d\omega=\int_a^b\nabla a\cdot\boldsymbol{r}'du=a(\boldsymbol{r}(b))-a(\boldsymbol{r}(a))$$

$$\int_{\partial\boldsymbol{r}}\omega=(-1)^{(1+1)}a(\boldsymbol{r}(b))-(-1)^{(1+1)}a(\boldsymbol{r}(a))$$

which is the same thing. This is the case of a conservative vector field, the potential function being a.

18.8.3 Green's Theorem

Next ω be a 1 form and let $p=k=2$ so $d\omega$ is a 2 form and ω is a 1 form. Say

$$\omega=P(x,y)\,dx+Q(x,y)\,dy$$

$$d\omega=P_y(x,y)\,dy\wedge dx+Q_x dx\wedge dy$$

Recall that terms like $dx \wedge dx$ are zero because they result in a determinant which equals 0. Then let $\boldsymbol{r}$ be smooth and map $[a,b] \times [c,d]$ to $\mathbb{R}^2$.

$$\boldsymbol{r}(s,u) \equiv (x(s,u), y(s,u))^T.$$

Then $\int_{\boldsymbol{r}} d\omega \equiv$

$$\int_a^b \int_c^d P_y(x(s,u), y(s,u)) \begin{vmatrix} y_s & y_u \\ x_s & x_u \end{vmatrix} + Q_x(x(s,u), y(s,u)) \begin{vmatrix} x_s & x_u \\ y_s & y_u \end{vmatrix} dsdu$$

$$= \int_a^b \int_c^d [Q_x(x(s,u), y(s,u)) - P_y(x(s,u), y(s,u))] (x_s y_u - x_u y_s) dsdu$$

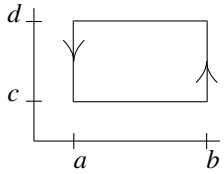
Let $U = \boldsymbol{r}([a,b] \times [c,d])$. Then by the change of variables theorem, this equals

$$\int_U (Q_x(x,y) - P_y(x,y)) \operatorname{sgn}(x_s y_u - x_u y_s) dxdy$$

Next consider $\int_{\partial \boldsymbol{r}} \omega$. For $\boldsymbol{r}(s,u) = (x(s,u), y(s,u))^T$, this equals

$$\begin{aligned}
&\int_c^d P(x(1,u), y(1,u)) \frac{\partial x(1,u)}{\partial u} + Q(x(1,u), y(1,u)) \frac{\partial y(1,u)}{\partial u} du \\
&- \int_c^d P(x(0,u), y(0,u)) \frac{\partial x(0,u)}{\partial u} + Q(x(0,u), y(0,u)) \frac{\partial y(0,u)}{\partial u} du \\
&+ \int_a^b P(x(s,0), y(s,0)) \frac{\partial x(s,0)}{\partial s} + Q(x(s,0), y(s,0)) \frac{\partial y(s,0)}{\partial s} ds \\
&- \int_a^b P(x(s,1), y(s,1)) \frac{\partial x(s,1)}{\partial s} + Q(x(s,1), y(s,1)) \frac{\partial y(s,1)}{\partial s} ds
\end{aligned}$$

This is computing a line integral by summing the contributions around the edges of $[a,b] \times [c,d]$. It is $\int_C P dx + Q dy$ where the orientation on this curve C comes from the counter clockwise orientation on the boundary of $[a,b] \times [c,d]$ as shown in the picture.



Thus if $x_s y_u - x_u y_s > 0$, this is just Green's theorem from calculus. Thus this gives a proof of this important theorem provided the region U is the C^1 image of a rectangle. One could generalize to consider chains of rectangles which, as mentioned above yields fairly general surfaces in $\mathbb{R}^2$. One could also include Lipschitz maps for even more generality. However, this falls short of the best results for Green's theorem which involve a rectifiable simple closed curve with U its interior. Rectifiable only requires the curve to have finite length.

18.8.4 Stoke's Theorem from Calculus

Next let $k = 2$ and $p = 3$. Thus $d\omega$ is a 2 form so ω is a 1 form. Say

$$\omega = P(x,y,z) dx + Q(x,y,z) dy + R(x,y,z) dz$$

Then

$$d\omega = P_y dy\wedge dx + P_z dz\wedge dx + Q_x dx\wedge dy + Q_z dz\wedge dy + R_x dx\wedge dz + R_y dy\wedge dz$$

Now let $\boldsymbol{r}:[0,1]^2\to\mathbb{R}^3, \boldsymbol{r}(s,u) = (x(s,u), y(s,u), z(s,u))$. Then letting the various functions be defined at $\boldsymbol{r}(s,u)$,

$$\int_{\boldsymbol{r}} d\omega = \int_{[0,1]^2} \left(P_y \det\begin{pmatrix} y_s & y_u \\ x_s & x_u \end{pmatrix} + P_z \det\begin{pmatrix} z_s & z_u \\ x_s & x_u \end{pmatrix} + Q_x \det\begin{pmatrix} x_s & x_u \\ y_s & y_u \end{pmatrix} \right.$$
$$\left. + Q_z \det\begin{pmatrix} z_s & z_u \\ y_s & y_u \end{pmatrix} + R_x \det\begin{pmatrix} x_s & x_u \\ z_s & z_u \end{pmatrix} + R_y \det\begin{pmatrix} y_s & y_u \\ z_s & z_u \end{pmatrix} \right) dsdu$$

This equals

$$\int_{[0,1]^2} (R_y - Q_z)(y_s z_u - y_u z_s) + (P_z - R_x)(x_u z_s - x_s z_u)$$
$$+ (Q_x - P_y)(x_s y_u - x_u y_s)\, dsdu$$

By the definition of surface area on $S \equiv \boldsymbol{r}\left([0,1]^2\right)$, see Definition 14.3.1 the area increment on the surface $\boldsymbol{r}\left([0,1]^2\right)$ is

$$\begin{aligned} & \sqrt{|y_s z_u - y_u z_s|^2 + |x_u z_s - x_s z_u|^2 + |x_s y_u - x_u y_s|^2}\, dsdu \\ = \quad & \det\left(D\boldsymbol{r}(u,s)^* D\boldsymbol{r}(s,u)\right)^{1/2} dsdu \end{aligned}$$

also the vector $(R_y - Q_z)\boldsymbol{i} + (P_z - R_x)\boldsymbol{j} + (Q_x - P_y)\boldsymbol{k}$ is the curl of the vector field $\boldsymbol{F} \equiv P\boldsymbol{i} + Q\boldsymbol{j} + R\boldsymbol{j}$. Thus in more familiar calculus notation, the above integral is of the form $\int_{S\equiv\boldsymbol{r}([0,1]^2)} \nabla\times\boldsymbol{F}\cdot\boldsymbol{p}\, dS$ where $\boldsymbol{p}$ is a vector which is in the direction of

$$(y_s z_u - y_u z_s)\boldsymbol{i} + (x_u z_s - x_s z_u)\boldsymbol{j} + (x_s y_u - x_u y_s)\boldsymbol{k}.$$

It happens that this vector $\boldsymbol{p}$ is perpendicular to the surface S at every point where $\boldsymbol{r}_s\times\boldsymbol{r}_u \neq \boldsymbol{0}$,which is seen to occur in the above formula. Recall that this follows from noting that, as (hopefully) discussed in beginning calculus,you have $\boldsymbol{r}_s\cdot\boldsymbol{r}_s\times\boldsymbol{r}_u = \boldsymbol{r}_u\cdot\boldsymbol{r}_s\times\boldsymbol{r}_u = 0$ and this is sufficient to claim, based on geometric reasoning that it is perpendicular to the surface. Thus $\int_{\boldsymbol{r}} d\omega$ is one side of the usual Stoke's theorem from calculus. You could generalize to chains of rectangles as well.

Next consider what happens with $\int_{\partial\boldsymbol{r}}\omega$. This is just like it was with Green's theorem but with more terms. To save on space, $P(x(1,u), y(1,u), z(1,u))$ is denoted as $P(1,u)$ with similar considerations for Q and R. Then this results in

$$\int_0^1 \left(P(1,u)\frac{\partial x(1,u)}{\partial u} + Q(1,u)\frac{\partial y(1,u)}{\partial u} + R(1,u)\frac{dz(1,u)}{du} \right) du$$
$$- \int_0^1 \left(P(0,u)\frac{\partial x(0,u)}{\partial u} + Q(0,u)\frac{\partial y(0,u)}{\partial u} + R(1,u)\frac{dz(1,u)}{du} \right) du$$
$$+ \int_0^1 P(s,0)\frac{\partial x(s,0)}{\partial s} + Q(s,0)\frac{\partial y(s,0)}{\partial s} + R(s,0)\frac{dz(s,0)}{ds} ds$$
$$- \int_0^1 P(s,1)\frac{\partial x(s,1)}{\partial s} + Q(s,1)\frac{\partial y(s,1)}{\partial s} + R(s,1)\frac{dz(s,1)}{ds} ds$$

As before, this is a line integral of the form $\int_C Pdx+Qdy+Rdz$. where C is an oriented curve bounding the surface S. This orientation will determine the direction of the vector $\boldsymbol{p}$ above.

18.8.5 The Divergence Theorem

In this case, we have a parameter domain $[\boldsymbol{a},\boldsymbol{b}] \subseteq \mathbb{R}^p$ and the differential form is

$$\omega = \sum_{r=1}^{p} \alpha_r(\boldsymbol{x})(-1)^{r-1} dx_1 \wedge \cdots \wedge d\hat{x}_r \wedge \cdots \wedge dx_p$$

where $d\hat{x}_r$ with the hat means that dx_r is omitted. The reason for the $(-1)^{r-1}$ is to make minus signs disappear in $d\omega$. This led to the divergence theorem

$$\pm \int_{\boldsymbol{r}([\boldsymbol{a},\boldsymbol{b}])} \sum_{j=1}^{p} \frac{\partial \alpha_j}{\partial x_j}(\boldsymbol{x})\, d\mathscr{H}^k = \int_{\boldsymbol{r}(\partial[\boldsymbol{a},\boldsymbol{b}])} \sum_{j=1}^{p} \alpha_j(\boldsymbol{x}) N^j(\boldsymbol{x})\, d\mathscr{H}^{k-1}(\boldsymbol{x})$$

where the $+$ sign holds if and only if $\frac{\partial(x_1,\cdots,x_p)}{\partial(u_1,\cdots,u_p)} \geq 0$. We needed to have $\boldsymbol{r}$ is one to one off a set $\hat{S}$ where $\boldsymbol{r}(\hat{S})$ has $\mathscr{H}^p$ measure zero and $\boldsymbol{r}_{b_l}, \boldsymbol{r}_{a_l}$ are also one to one or at least one to one off a set S where $\boldsymbol{r}_{b_l}(S), \boldsymbol{r}_{a_l}(S)$ have $\mathscr{H}^{p-1}$ measure zero. It remains to consider the vector $\boldsymbol{N}$ which has j^{th} component N^j. I want to argue that this vector is a.e. normal to $\boldsymbol{r}(\partial[\boldsymbol{a},\boldsymbol{b}])$ and that sometimes it is an outer normal. Recall that for $\boldsymbol{x} = \boldsymbol{r}_{b_l}(\boldsymbol{u}_l)$, $N^j(\boldsymbol{x}) = (-1)^{j+l} \frac{\partial(x_1,\cdots,\hat{x}_j,\cdots,x_p)}{\partial(u_1,\cdots,\hat{u}_l\cdots,u_p)} \frac{1}{J_*(\boldsymbol{x})}$. Thus for a point on the boundary, for $i \neq l$,

$$\boldsymbol{r}_{u_i} \cdot \boldsymbol{N} = \sum_{j=1}^{p} x_{j,u_i} N^j(\boldsymbol{x}) = \sum_{j=1}^{p} x_{j,u_i} (-1)^{l+j} \frac{\partial(x_1,\cdots,\hat{x}_j,\cdots,x_p)}{\partial(u_1,\cdots,\hat{u}_l\cdots,u_p)} \frac{1}{J_*(\boldsymbol{x})}$$

$$= \frac{1}{J_*(\boldsymbol{x})} \det\begin{pmatrix} \boldsymbol{x}_{u_1} & \cdots & \boldsymbol{x}_{u_i} & \cdots & \boldsymbol{x}_{u_i} & \cdots & \boldsymbol{x}_{u_p} \end{pmatrix} = 0$$

since it is a determinant with two equal columns. This involved expanding along the l^{th} column which was filled by $\boldsymbol{x}_{u_i}$. Also, expanding along this column,

$$\boldsymbol{r}_{u_l} \cdot \boldsymbol{N} = \frac{1}{J_*(\boldsymbol{x})} \det(D\boldsymbol{r}) = \operatorname{sgn}(\det(D\boldsymbol{r}))$$

It follows that $\boldsymbol{N}$ is perpendicular to $\boldsymbol{r}(\partial[\boldsymbol{a},\boldsymbol{b}])$ and that the angle between $\boldsymbol{r}_{u_l}$ and $\boldsymbol{N}$ is no more than 90 degrees if $\frac{\partial(x_1,\cdots,x_p)}{\partial(u_1,\cdots,u_p)} \geq 0$. This yields the following.

Theorem 18.8.1 *Let $\boldsymbol{r}$ be Lipschitz and one to one off S where $\boldsymbol{r}(S)$ has $\mathscr{H}^p$ measure zero and suppose a similar condition holds for $\boldsymbol{r}_{a_l}$ and $\boldsymbol{r}_{b_l}$. Let α_j be C^1*

$$\omega = \sum_{j=1}^{p} \alpha_j(\boldsymbol{x})(-1)^{j-1} dx_1 \wedge \cdots \wedge d\hat{x}_j \wedge \cdots \wedge dx_p,\ \boldsymbol{\alpha}(\boldsymbol{x}) = (\alpha_1(\boldsymbol{x}), ..., \alpha_p(\boldsymbol{x}))$$

and suppose $\det(D\boldsymbol{r}(\boldsymbol{u})) \geq 0$. *Then*

$$\int_{\boldsymbol{r}} d\omega = \int_{\boldsymbol{r}([\boldsymbol{a},\boldsymbol{b}])} \sum_{j} \frac{\partial \alpha_j}{\partial x_j}(\boldsymbol{x})\, d\mathscr{H}^p = \int_{\boldsymbol{r}(\partial[\boldsymbol{a},\boldsymbol{b}])} \boldsymbol{\alpha} \cdot \boldsymbol{N} d\mathscr{H}^{p-1} \tag{18.10}$$

where $\boldsymbol{N}$ *is an outer normal in the sense that the angle between the vector* $\boldsymbol{r}_{u_l}$ *and* $\boldsymbol{N}$ *is no more than 90 degrees if* $\frac{\partial(x_1,\cdots,x_p)}{\partial(u_1,\cdots,u_p)} \geq 0$.

Also, if you know the divergence theorem, then you can directly give the usual Calculus version of Green's and Stoke's theorems from Calculus. This is developed in the exercises.

18.9 Exercises

1. Let $f:\mathbb{R}^n \to \mathbb{R}$ be given by $f(\boldsymbol{x}) \equiv \left(\sum_{i=1}^n \frac{x_i^2}{a_i^2}\right)^{1/2}$ where each $a_i > 0$. Thus for $y > 0$ $f^{-1}(y)$ is the boundary of an n dimensional ellipsoid. Using change of variables formula and the coarea formula, find the area of $f^{-1}(r)$.

2. Let $\boldsymbol{\pi}:\mathbb{R}^n \to \mathbb{R}^m$ be defined as $\boldsymbol{\pi}(\boldsymbol{x}) = \boldsymbol{x_i}$ where $\boldsymbol{i} = (i_1, i_2, ..., i_m)$. What does Theorem 17.9.1 say if $\boldsymbol{\pi} = \boldsymbol{f}$ in this theorem?

3. In calculus, you found the area of the parallelogram determined by two vectors $\boldsymbol{u}, \boldsymbol{v}$ in $\mathbb{R}^3$ by taking the magnitude $|\boldsymbol{u} \times \boldsymbol{v}|$, meaning the Euclidean norm of the cross product. Show that you get the same answer by forming

$$\left(\det\begin{pmatrix} \boldsymbol{u} & \boldsymbol{v} \end{pmatrix}^* \begin{pmatrix} \boldsymbol{u} & \boldsymbol{v} \end{pmatrix}\right)^{1/2}$$

   where here you have $\begin{pmatrix} \boldsymbol{u} & \boldsymbol{v} \end{pmatrix}$ is the matrix which has $\boldsymbol{u}$ as first column and $\boldsymbol{v}$ as second column.

4. In calculus, you also found the volume of a parallelepiped determined by the vectors $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}$ by $|\boldsymbol{u} \cdot (\boldsymbol{v} \times \boldsymbol{w})|$, the absolute value of the box product. Show this turns out to be the same thing as

$$\left(\det\begin{pmatrix} \boldsymbol{u} & \boldsymbol{v} & \boldsymbol{w} \end{pmatrix}^* \begin{pmatrix} \boldsymbol{u} & \boldsymbol{v} & \boldsymbol{w} \end{pmatrix}\right)^{1/2}.$$

5. Imagine a fluid which does not move. Let $B(\boldsymbol{x}, \varepsilon)$ be a small ball in this fluid. Use the Euclidean norm. Then the force exerted on the ball of fluid is $-\int_{\partial B(\boldsymbol{x},\varepsilon)} p\boldsymbol{n} dA$ where p is the pressure. Here $\boldsymbol{n}$ is the unit exterior normal. Now the force acting on the ball from gravity is $-g\boldsymbol{k}\int_{B(\boldsymbol{x},\varepsilon)} \rho dV$ where ρ signifies the density of the fluid and $\boldsymbol{k}$ signifies the direction which is up. The vectors $\boldsymbol{i}, \boldsymbol{j}$ are in the direction of the positive x and y axes respectively. These two forces add to $\boldsymbol{0}$ because it is given that the fluid does not move. Use the divergence theorem to show that $\nabla p = \rho g \boldsymbol{k}$. This is a really neat result.

6. Archimedes principle states that when a solid body is immersed in a static fluid, the force acting on the body by the fluid is directly up and equals the total weight of the fluid displaced. Surely this is an amazing result. It doesn't matter about the shape of the body. Remember that the weight is the acceleration of gravity times the mass. **Hint:** You need to start with the force acting on the body B by the fluid which is $-\int_{\partial B} p\boldsymbol{n} dA$. Assume the divergence theorem holds for B. As shown, this is typically the case.

7. You have a closed region R which is **fixed in space.** A fluid is flowing through R. The density of this fluid is $\rho(t,\boldsymbol{x})$ where $\boldsymbol{x}$ gives the coordinates in space and ρ depends on t because it might change as time progresses. The velocity of this fluid is $\boldsymbol{v}(t,\boldsymbol{x})$. Then the rate at which the fluid crosses a surface S from one side to the other is $\int_S \rho \boldsymbol{v}\cdot\boldsymbol{n}dS$ where $\boldsymbol{n}$ is the unit normal to the surface which points in the direction of interest. You can think about this a little and see that it is a reasonable claim. If, for example, $\boldsymbol{v}\cdot\boldsymbol{n}=0$, then the velocity of the fluid would be parallel to the surface so it would not cross it at all. Also, the total mass of the fluid which is in R is $\int_R \rho dV$. Assuming anything you like about regularity, which is what we do in situations like this, explain using the divergence theorem why $\int_V \frac{\partial\rho}{\partial t}+\nabla\cdot(\rho\boldsymbol{v})\,dV = 0$. Recall that $\nabla\cdot\boldsymbol{F}$ denotes the divergence of $\boldsymbol{F}$. Now explain why it should be the case that $\frac{\partial\rho}{\partial t}+\nabla\cdot(\rho\boldsymbol{v})=0$. This is called the balance of mass equation.

8. The permutation symbol is $\varepsilon_{ijk}=1$ if the permutation $\begin{pmatrix}1&2&3\\ i&j&k\end{pmatrix}$ is even and -1 if this permutation is odd. Such a permutation is odd or even depending on whether it requires an odd or even number of switches to obtain $(1,2,3)$. Thus $\varepsilon_{123}=1,\varepsilon_{213}=-1$ and so forth. Show that $\sum_k \varepsilon_{ijk}\varepsilon_{irs}=\delta_{jr}\delta_{ks}-\delta_{js}\delta_{kr}$. We often agree to add over a repeated index to avoid having to write the summation symbol. Thus $\sum_k \varepsilon_{ijk}\varepsilon_{irs}=\varepsilon_{ijk}\varepsilon_{irs}$. If you do this, avoid having the repeated index repeated more than once in any term. Here δ_{ij} is 1 if $i=j$ and 0 if $i\neq j$.

9. Let U be a bounded open set in $\mathbb{R}^p$ and suppose $u\in C^2(U)\cap C\left(\overline{U}\right)$ such that $\nabla^2 u\geq 0$ in U. Then letting $\partial U=\overline{U}\setminus U$, it follows that

$$\max\left\{u(\boldsymbol{x}):\boldsymbol{x}\in\overline{U}\right\}=\max\left\{u(\boldsymbol{x}):\boldsymbol{x}\in\partial U\right\}.$$

The symbol ∇^2 is the Laplacian. Thus $\nabla^2 u=\sum_i u_{x_ix_i}$. In terms of repeated index summation convention, $\nabla^2 u=u_{,ii}$. **Hint:** Suppose this does not happen. Then there exists $\boldsymbol{x}_0\in U$ with

$$u(\boldsymbol{x}_0)>\max\left\{u(\boldsymbol{x}):\boldsymbol{x}\in\partial U\right\}.$$

Since U is bounded, there exists $\varepsilon>0$ such that

$$u(\boldsymbol{x}_0)>\max\left\{u(\boldsymbol{x})+\varepsilon|\boldsymbol{x}|^2:\boldsymbol{x}\in\partial U\right\}.$$

Therefore, $u(\boldsymbol{x})+\varepsilon|\boldsymbol{x}|^2$ also has its maximum in U because for ε small enough,

$$u(\boldsymbol{x}_0)+\varepsilon|\boldsymbol{x}_0|^2>u(\boldsymbol{x}_0)>\max\left\{u(\boldsymbol{x})+\varepsilon|\boldsymbol{x}|^2:\boldsymbol{x}\in\partial U\right\}$$

for all $\boldsymbol{x}\in\partial U$. Now let $\boldsymbol{x}_1$ be the point in U where $u(\boldsymbol{x})+\varepsilon|\boldsymbol{x}|^2$ achieves its maximum. Now recall the second derivative test from single variable calculus. Explain why at a local maximum of f you must have $\nabla^2 f\leq 0$. Apply this to the function $\boldsymbol{x}\to u(\boldsymbol{x})+\varepsilon|\boldsymbol{x}|^2$ at the point $\boldsymbol{x}_1$ and get a contradiction. This is called the weak maximum principle.

10. Review the cross product from calculus. Show that in $\mathbb{R}^3$, $(\boldsymbol{a}\times\boldsymbol{b})_i=\varepsilon_{ijk}a_jb_k$ where summation is over repeated indices. Using the above reduction identity of Problem 8, simplify $(\boldsymbol{a}\times\boldsymbol{b})\times\boldsymbol{c}$ in terms of dot products. Then do the same for $\boldsymbol{a}\times(\boldsymbol{b}\times\boldsymbol{c})$.

11. Show that $\nabla \cdot (\nabla \times \boldsymbol{v}) = 0$. Now show $\int_{\partial V} \nabla \times \boldsymbol{v} \cdot \boldsymbol{n}\, dA = 0$ where V is a region for which the divergence theorem holds and $\boldsymbol{v}$ is a C^2 vector field. $\nabla \times \boldsymbol{v}$ is the curl of $\boldsymbol{v}$. In the new notation, $(\nabla \times \boldsymbol{v})_i = \varepsilon_{ijk}\partial_j v_k$ where ∂_j is an operator which means to take the partial derivative with respect to x_j. It is understood here that the coordinates are rectangular coordinates. The first part of this is real easy if you remember the big theorem about equality of mixed partial derivatives.

12. Let U be a bounded open set in $\mathbb{R}^2$ which has a Lipschitz boundary so that the divergence theorem holds. Let the axes be oriented in the usual way as in calculus. Let $P(x,y), Q(x,y)$ be two smooth functions defined on $\bar{U}$. What does the divergence theorem say for the vector field $(Q,-P)$? If $\boldsymbol{r} : [a,b] \to \mathbb{R}^2$ is Lipschitz with $\boldsymbol{r}(a) = \boldsymbol{r}(b)$ and $\boldsymbol{r}$ is one to one with $\partial U = \boldsymbol{r}([a,b])$ and $\boldsymbol{r}'(t)$ is in the direction of $\boldsymbol{k} \times \boldsymbol{n}$ so $\boldsymbol{r}'(t) \times \boldsymbol{k}$ is in the direction of $\boldsymbol{n}$, a statement about orientation. Thus for $\boldsymbol{r}(t) = (x(t), y(t))$,

$$\boldsymbol{n} = c \begin{vmatrix} \boldsymbol{i} & \boldsymbol{j} & \boldsymbol{k} \\ x'(t) & y'(t) & 0 \\ 0 & 0 & 1 \end{vmatrix} = c\left(y'(t)\,\boldsymbol{i} - x'(t)\,\boldsymbol{j}\right), \text{ so } c = \frac{1}{\sqrt{y'(t)^2 + x'(t)^2}}$$

for some c a positive constant. (This is the cross product from calculus. Review this.), use the area formula to describe the boundary integral from the divergence theorem as an integral over $[a,b]$. We write this integral as $\int_{\partial U} P\,dx + Q\,dy$. This yields a version of Green's theorem, $\int_U Q_x - P_y\, dm_2 = \int_{\partial U} P\,dx + Q\,dy$ provided ∂U is a Lipschitz curve oriented as just described. However, more generality is possible, although I am not sure how far this has been generalized. You really only need to have the boundary of U be a rectifiable simple closed curve meaning it has finite length. In this setting, the Jordan curve theorem makes it possible to make sense of Green's theorem and in fact it holds. In what was just discussed, $\bar{U}$ was the Lipschitz image of some rectangle. The general version only requires that the boundary of U be the image of the unit circle. Nevertheless, this version in this problem is pretty good.

13. Let U be a bounded open set for which Green's theorem holds. Let C be the oriented boundary consistent with Green's theorem. Show: area of $U = \int_C x\,dy$.

14. Let U be an open set with which is on one side of its boundary as above with the boundary being the image of a Lipschitz map which is one to one. (Note this condition eliminates the curve crossing itself.) Now suppose you have a closed polygonal curve going from $(x_0,y_0) \to (x_1,y_1) \to (x_2,y_2) \cdots (x_p,y_p) = (x_0,y_0)$ oriented such that the direction of motion is in the direction $\boldsymbol{k} \times \boldsymbol{n}$ where $\boldsymbol{n}$ is the unit outer normal from the divergence theorem. Show that if U is this enclosed polygon, then

$$\text{Area of } U = \frac{1}{2}\sum_{k=1}^{n} (x_k + x_{k-1})(y_k - y_{k-1})$$

This is a pretty remarkable result. Just draw a few such polygons and think how you would find their area without it.

15. Orient the u,v axes just like the usual arrangement of the x and y axes, u axis like the x axis. Let $\boldsymbol{r} : U \to \mathbb{R}^3$ where U is an open subset of $\mathbb{R}^2$. Suppose $\boldsymbol{r}$ is C^2 and let

$\boldsymbol{F}$ be a C^1 vector field defined in V, an open set containing $\boldsymbol{r}(U)$. Show, using the above reduction identity of Problem 8 that

$$(\boldsymbol{r}_u \times \boldsymbol{r}_v) \cdot (\nabla \times \boldsymbol{F})(\boldsymbol{r}(u,v)) = ((\boldsymbol{F} \circ \boldsymbol{r})_u \cdot \boldsymbol{r}_v - (\boldsymbol{F} \circ \boldsymbol{r})_v \cdot \boldsymbol{r}_u)(u,v). \tag{18.11}$$

The left side is the dot product of the curl of the vector field $\boldsymbol{F}$ with a normal vector to the surface $\boldsymbol{r}(U)$, namely $\boldsymbol{r}_u \times \boldsymbol{r}_v$. Show that the right side can be written as $((\boldsymbol{F} \circ \boldsymbol{r}) \cdot \boldsymbol{r}_v)_u - ((\boldsymbol{F} \circ \boldsymbol{r}) \cdot \boldsymbol{r}_u)_v$ thanks to equality of mixed partial derivatives. Now suppose U is a region for which Green's theorem holds, the curve C bounding U being Lipschitz. Verify Stokes' theorem

$$\int_U (\boldsymbol{r}_u \times \boldsymbol{r}_v) \cdot (\nabla \times \boldsymbol{F})(\boldsymbol{r}(u,v))\, du dv = \int_{\boldsymbol{R}} (\boldsymbol{F} \circ \boldsymbol{r}) \cdot \boldsymbol{r}_u du + (\boldsymbol{F} \circ \boldsymbol{r}) \cdot \boldsymbol{r}_v dv$$

where $\boldsymbol{R}(t) = (u(t), v(t))$ for $t \to (u(t), v(t))$ being a parametrization of C oriented so that $(u'(t), v'(t), 0) \times \boldsymbol{k}$ is an exterior normal to U. Show the left integral can be written as $\int_{\boldsymbol{r}(U)} \nabla \times \boldsymbol{F} \cdot \boldsymbol{N} d\mathscr{H}^2$ where $\boldsymbol{N}$ is a unit normal to the surface $\boldsymbol{r}(U)$. Show the integral on the right can be written as the differential form $\int_{\boldsymbol{R}} F_1 dx + F_2 dy + F_3 dz$ where $\boldsymbol{R}(t) = \boldsymbol{r}(u(t), v(t))$.

16. Let $f : \mathbb{C} \to \mathbb{C}$. Thus if $x + iy \in \mathbb{C}$, $f(x+iy) = u(x,y) + iv(x,y)$ where u is the real part and v is the imaginary part. As in one variable calculus, we define

$$\lim_{h \to 0} \frac{f(z+h) - f(z)}{h} = f'(z)$$

and we say that this derivative exists exactly when this limit exists. Consider $h = it$ and then let $h = t$. Take limits in these two ways and conclude that if $f'(z)$ exists, then it is given by

$$f'(z) = u_x + iv_x = v_y - iu_y$$

Thus you have the Cauchy Riemann equations $u_x = v_y, v_x = -u_y$. Show that if u, v are both C^1, and these Cauchy Riemann equation hold, then the function will be differentiable. When this happens, we say the function is analytic. (In fact it can be shown that if the limit of the difference quotient exists, then these real and imaginary parts will automatically be continuous and the function will be analytic.)

17. For a function $f : \mathbb{C} \to \mathbb{C}$ which is continuous and $\gamma : [a,b] \to \Gamma \subseteq \mathbb{C}$ where Γ is a piecewise smooth curve, we define the contour integral

$$\int_\Gamma f(z)\, dz = \int_a^b f(\gamma(t))\, \gamma'(t)\, dt.$$

Show that this equals $F(\gamma(b)) - F(\gamma(a))$ if f is analytic, this for some function F. In particular, if Γ is a suitable closed curve, then $\int_\Gamma f(z)\, dz = 0$. This is Cauchy's theorem from complex analysis.

18. Suppose $f \in L^1(U)$ where U is some open set in $\mathbb{R}^p$. Go ahead and assume f is Borel measurable although it should work with f only Lebesgue measurable. Show there is a set of m_{p-1} measure zero N such that if $\boldsymbol{x}_p \equiv (x_1, x_2, \cdots, x_{p-1}) \notin N$, then $x_p \to f(\boldsymbol{x}_p, x_p)$ is in $L^1(U_{\boldsymbol{x}_p})$ where $U_{\boldsymbol{x}_p} = \{t : (\boldsymbol{x}_p, t) \in U\}$.

19. If $f \in L^1(U)$ and $f_{x_p} \in L^1(U)$ where $U \subseteq \mathbb{R}^p$ is a box like $\prod_k (a_k, b_k)$. Let f_{x_p} refer to the weak partial derivative. Can you show that for fixed $s, t \in (a_k, b_k)$ then for a.e. $\boldsymbol{x}_p$, $f(\boldsymbol{x}_p, t) - f(\boldsymbol{x}_p, s) = \int_s^t f_{x_p}(\boldsymbol{x}_p, \tau)\, d\tau$? Assume f is Borel measurable.

Part III

Abstract Theory

Chapter 19

Hausdorff Spaces and Measures

19.1 General Topological Spaces

It turns out that metric spaces are not sufficiently general for some applications. This section is a brief introduction to general topology. In making this generalization, the properties of balls in a metric space are stated as axioms for a subset of the power set of a given set X. This subset of the power set $\mathscr{P}(X)$ (set of all subsets) will be known as a basis for a topology. The properties of balls which are of importance are that the intersection of finitely many is the union of balls and that the union of all of them give the whole space. Recall that with a metric space, an open set was just one in which every point was an interior point. This simply meant that every point is contained in a ball which is contained in the given set. All that is being done here is to make these simple properties into axioms.

Definition 19.1.1 *Let X be a nonempty set and suppose $\mathscr{B} \subseteq \mathscr{P}(X)$. Then $\mathscr{B}$ is a basis for a topology if it satisfies the following axioms.*

1.) Whenever $p \in A \cap B$ for $A, B \in \mathscr{B}$, it follows there exists $C \in \mathscr{B}$ such that $p \in C \subseteq A \cap B$.

2.) $\cup \mathscr{B} = X$.

Then a subset U, of X is an open set if for every point $x \in U$, there exists $B \in \mathscr{B}$ such that $x \in B \subseteq U$. Thus the open sets are exactly those which can be obtained as a union of sets of $\mathscr{B}$. Denote these subsets of X by the symbol τ and refer to τ as the topology or the set of open sets.

Note that this is simply the analog of saying a set is open exactly when every point is an interior point.

Proposition 19.1.2 *Let X be a set and let $\mathscr{B}$ be a basis for a topology as defined above and let τ be the set of open sets determined by $\mathscr{B}$. Then*

$$\emptyset \in \tau,\ X \in \tau, \tag{19.1}$$

$$\text{If } \mathscr{C} \subseteq \tau, \text{ then } \cup \mathscr{C} \in \tau \tag{19.2}$$

$$\text{If } A, B \in \tau, \text{ then } A \cap B \in \tau. \tag{19.3}$$

Proof: If $p \in \emptyset$ then there exists $B \in \mathscr{B}$ such that $p \in B \subseteq \emptyset$ because there are no points in $\emptyset$. Therefore, $\emptyset \in \tau$. Now if $p \in X$, then by part 2.) of Definition 19.1.1 $p \in B \subseteq X$ for some $B \in \mathscr{B}$ and so $X \in \tau$.

If $\mathscr{C} \subseteq \tau$, and if $p \in \cup \mathscr{C}$, then there exists a set, $B \in \mathscr{C}$ such that $p \in B$. However, B is itself a union of sets from $\mathscr{B}$ and so there exists $C \in \mathscr{B}$ such that $p \in C \subseteq B \subseteq \cup \mathscr{C}$. This verifies 19.2.

Finally, if $A, B \in \tau$ and $p \in A \cap B$, then since A and B are themselves unions of sets of $\mathscr{B}$, it follows there exists $A_1, B_1 \in \mathscr{B}$ such that $A_1 \subseteq A, B_1 \subseteq B$, and $p \in A_1 \cap B_1$. Therefore, by 1.) of Definition 19.1.1 there exists $C \in \mathscr{B}$ such that $p \in C \subseteq A_1 \cap B_1 \subseteq A \cap B$, showing that $A \cap B \in \tau$ as claimed. Of course from the above, if $A \cap B = \emptyset$, then $A \cap B \in \tau$. ■

Definition 19.1.3 *A set X together with such a collection of its subsets satisfying 19.1-19.3 is called a topological space. τ is called the topology or set of open sets of X.*

Definition 19.1.4 *A topological space is said to be Hausdorff if whenever p and q are distinct points of X, there exist disjoint open sets U, V such that $p \in U$, $q \in V$. In other words points can be separated with open sets.*

Hausdorff

Definition 19.1.5 *A subset of a topological space is said to be closed if its complement is open. Let p be a point of X and let $E \subseteq X$. Then p is said to be a limit point of E if every open set containing p contains a point of E distinct from p.*

Theorem 19.1.6 *If (X, τ) is a Hausdorff space and if $p \in X$, then $\{p\}$ is a closed set.*

Proof: If $x \neq p$, there exist open sets U and V such that $x \in U, p \in V$ and $U \cap V = \emptyset$. Therefore, $\{p\}^C$ is an open set so $\{p\}$ is closed. ■

It would have been enough to assume that if $x \neq y$, then there exists an open set containing x which does not contain y.

Proposition 19.1.7 *If (X, τ) is a Hausdorff space then a point p is a limit point of a set E if and only if every open set containing p contains infinitely many points of E each different than p.*

Proof: $\Leftarrow$ is obvious. Consider $\Rightarrow$. If p is a limit point and if U is an open set containing p but there are only finitely many points of E different than p contained in $U, \{q_i\}_{i=1}^m$, then consider $V \equiv U \cap \cap_{i=1}^m \{q_i\}^C$ which is an open set because each $\{q_i\}^C$ is open. This is because if $x \neq q_i$ there exists open V_{q_i} containing x such that $q_i \notin V_{q_i}$ and so V is a finite intersection of open sets. Therefore, there is a $q_{m+1} \in V \setminus \{p\}$, a contradiction. ■

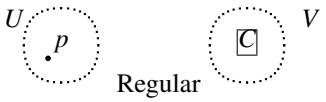
Theorem 19.1.8 *A subset E, of X is closed if and only if it contains all its limit points. A set is closed if and only if its complement is open and a set is open if and only if its complement is closed.*

Proof: Suppose first that E is closed and let x be a limit point of E. Is $x \in E$? If $x \notin E$, then E^C is an open set containing x which contains no points of E, a contradiction. Thus $x \in E$.

Now suppose E contains all its limit points. Is E^C open? If $x \in E^C$, then x is not a limit point of E because E has all its limit points and so there exists an open set, U containing x such that U contains no point of E other than x. Since $x \notin E$, it follows that $x \in U \subseteq E^C$ which implies E^C is an open set because this shows E^C is the union of open sets.

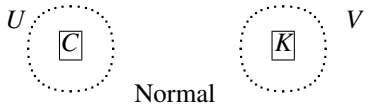
By definition, E closed $\Rightarrow E^C$ is open. If E^C is open, then no point of E^C can be a limit point of E and so E is closed since it contains all its limit points so E^C open $\Rightarrow E$ closed. ■

Definition 19.1.9 *A topological space (X, τ) is said to be regular if whenever C is a closed set and p is a point not in C, there exist disjoint open sets U and V such that $p \in U$, $C \subseteq V$. Thus a closed set can be separated from a point not in the closed set by two disjoint open sets.*



Regular

Definition 19.1.10 *The topological space, (X,τ) is said to be normal if whenever C and K are disjoint closed sets, there exist disjoint open sets U and V such that $C \subseteq U$, $K \subseteq V$. Thus any two disjoint closed sets can be separated with open sets.*



Normal

Definition 19.1.11 *Let E be a subset of X. $\overline{E}$ is defined to be the smallest closed set containing E.*

Lemma 19.1.12 *The above definition is well defined.*

Proof: Let $\mathscr{C}$ denote all the closed sets which contain E. Then $\mathscr{C}$ is nonempty because $X \in \mathscr{C}$.

$$(\cap\{A : A \in \mathscr{C}\})^C = \cup\left\{A^C : A \in \mathscr{C}\right\},$$

an open set which shows that $\cap\mathscr{C}$ is a closed set and is the smallest closed set which contains E.

Theorem 19.1.13 $\overline{E} = E \cup \{\text{limit points of } E\}$.

Proof: Let $x \in \overline{E}$ and suppose that $x \notin E$. If x is not a limit point either, then there exists an open set U,containing x which does not intersect E. But then U^C is a closed set which contains E which does not contain x, contrary to the definition that $\overline{E}$ is the intersection of all closed sets containing E. Therefore, x must be a limit point of E after all.

Now $E \subseteq \overline{E}$ so suppose x is a limit point of E. Is $x \in \overline{E}$? If H is a closed set containing E, which does not contain x, then H^C is an open set containing x which contains no points of E other than x negating the assumption that x is a limit point of E. ■

The following is the definition of continuity in terms of general topological spaces. It is really just a generalization of the $\varepsilon - \delta$ definition of continuity given in calculus.

Definition 19.1.14 *Let (X,τ) and (Y,η) be two topological spaces and let $f : X \to Y$. f is continuous at $x \in X$ if whenever V is an open set of Y containing $f(x)$, there exists an open set $U \in \tau$ such that $x \in U$ and $f(U) \subseteq V$. f is continuous if $f^{-1}(V) \in \tau$ whenever $V \in \eta$.*

Then the following comes from the definition.

Proposition 19.1.15 *In the situation of Definition 19.1.14 f is continuous if and only if f is continuous at every point of X.*

Proof: $\Rightarrow$ Suppose f is continuous and let $f(x) \in V$ an open set in Y. Then $x \in f^{-1}(V) \equiv U \in \tau$. $\Leftarrow$ Next suppose f is continuous at every point. Then if $V \in \eta$, and $x \in f^{-1}(V)$, continuity at x implies there is open $U_x \subseteq f^{-1}(V)$. Thus $f^{-1}(V) = \cup_{x \in f^{-1}(V)} U_x$ and so $f^{-1}(V)$ is open. ■

Definition 19.1.16 *Let (X_i,τ_i) be topological spaces. $\prod_{i=1}^n X_i$ is the Cartesian product. Define a product topology as follows. Let $\mathscr{B} = \prod_{i=1}^n A_i$ where $A_i \in \tau_i$. Then $\mathscr{B}$ is a basis for the product topology.*

Theorem 19.1.17 *The set $\mathscr{B}$ of Definition 19.1.16 is a basis for a topology as claimed.*

Proof: Suppose $\boldsymbol{x} \in (\prod_{i=1}^n A_i) \cap (\prod_{i=1}^n B_i)$ where A_i and B_i are open sets. Suppose that $\boldsymbol{x} = (x_1, \cdots, x_n)$. Then $x_i \in A_i \cap B_i$ for each i. Therefore, $\boldsymbol{x} \in \prod_{i=1}^n A_i \cap B_i \in \mathscr{B}$ and $\prod_{i=1}^n A_i \cap B_i \subseteq \prod_{i=1}^n A_i$. ■

The definition of compactness is also considered for a general topological space. This is given next.

Definition 19.1.18 *A subset, E, of a topological space (X, τ) is said to be compact if whenever $\mathscr{C} \subseteq \tau$ and $E \subseteq \cup\mathscr{C}$, there exists a finite subset of $\mathscr{C}$, $\{U_1 \cdots U_n\}$, such that $E \subseteq \cup_{i=1}^n U_i$. (Every open covering admits a finite subcovering.) E is precompact if $\overline{E}$ is compact. A topological space is called locally compact if it has a basis $\mathscr{B}$, with the property that $\overline{B}$ is compact for each $B \in \mathscr{B}$.*

In general topological spaces there may be no concept of "bounded". Even if there is, closed and bounded is not necessarily the same as compactness. However, in any Hausdorff space every compact set must be a closed set.

Theorem 19.1.19 *If (X, τ) is a Hausdorff space, then every compact subset must also be a closed set.*

Proof: Suppose $p \notin K$. For each $x \in X$, there exist open sets, U_x and V_x such that $x \in U_x$, $p \in V_x$, and $U_x \cap V_x = \emptyset$. If K is assumed to be compact, there are finitely many of these sets, $U_{x_1}, \cdots, U_{x_m}$ which cover K. Then let $V \equiv \cap_{i=1}^m V_{x_i}$. It follows that V is an open set containing p which has empty intersection with each of the U_{x_i}. Consequently, V contains no points of K and is therefore not a limit point of K. ■

Definition 19.1.20 *If every finite subset of a set P whose elements are sets has nonempty intersection, the set P is said to have the finite intersection property.*

Theorem 19.1.21 *Let $\mathscr{K}$ be a set whose elements are compact subsets of a Hausdorff topological space, (X, τ). Suppose $\mathscr{K}$ has the finite intersection property. Then $\emptyset \neq \cap\mathscr{K}$.*

Proof: Suppose to the contrary that $\emptyset = \cap\mathscr{K}$. Then consider $\mathscr{C} \equiv \left\{K^C : K \in \mathscr{K}\right\}$. It follows $\mathscr{C}$ is an open cover of K_0 where K_0 is any particular element of $\mathscr{K}$. But then there are finitely many $K \in \mathscr{K}$, $K_1, \cdots, K_r$ such that $K_0 \subseteq \cup_{i=1}^r K_i^C$ implying that

$$K_0 \cap (\cap_{i=1}^r K_i) \subseteq \left(\cup_{i=1}^r K_i^C\right) \cap (\cap_{i=1}^r K_i) = (\cap_{i=1}^r K_i)^C \cap (\cap_{i=1}^r K_i) = \emptyset,$$

contradicting the finite intersection property. ■

There is a fundamental theorem, called Urysohn's lemma which is valid for locally compact Hausdorff spaces which is presented next.

Lemma 19.1.22 *Let X be a locally compact Hausdorff space and let $K \subseteq V \subseteq X$ where K is compact and V is open. Then there exists an open set U_k containing k such that $\overline{U_k}$ is compact and $U_k \subseteq \overline{U}_k \subseteq V$. Also there exists U such that $\overline{U}$ is compact and $K \subseteq U \subseteq \overline{U} \subseteq V$.*

Proof: Since X is locally compact, there exists a basis of open sets whose closures are compact $\mathscr{U}$. Denote by $\mathscr{C}$ the set of all $U \in \mathscr{U}$ which contain k and let $\mathscr{C}'$ denote the set of all closures of these sets of $\mathscr{C}$ intersected with the closed set V^C. Thus $\mathscr{C}'$ is a collection of compact sets. There are finitely many of the sets of $\mathscr{C}'$ which have empty intersection. If not, then $\mathscr{C}'$ has the finite intersection property and so there exists a point p in all of them. Since X is a Hausdorff space, there exist disjoint basic open sets from $\mathscr{U}$, A, B such that $k \in A$ and $p \in B$. Therefore, $p \notin \overline{A}$ contrary to the above requirement that p be in all such sets. It follows there are sets $A_1, \cdots, A_m$ in $\mathscr{C}$ such that $V^C \cap \overline{A_1} \cap \cdots \cap \overline{A_m} = \emptyset$.Let $U_k \equiv A_1 \cap \cdots \cap A_m$. Then $\overline{U_k} \subseteq \overline{A_1} \cap \cdots \cap \overline{A_m}$ and so it has empty intersection with V^C. Thus it is contained in V. Also $\overline{U_k}$ is a closed subset of the compact set $\overline{A_1}$ so it is compact and $k \in U_k$.

For the second part, consider all such U_k. Since K is compact, there are finitely many which cover K $U_{k_1}, \cdots, U_{k_n}$. Then let $U \equiv \cup_{i=1}^n U_{k_i}$. It follows that $\overline{U} = \cup_{i=1}^n \overline{U_{k_i}}$ and each of these is compact so this set works. ■

The following is Urysohn's lemma for locally compact Hausdorff spaces.

Theorem 19.1.23 *(Urysohn) Let (X, τ) be locally compact and let $H \subseteq U$ where H is compact and U is open. Then there exists $g : X \to [0,1]$ such that g is continuous, $g(x) = 1$ on H and $g(x) = 0$ if $x \notin V$ for some open set V such that $\overline{V} \subseteq U$ such that $\overline{V}$ is compact.*

Proof: This involves using Lemma 19.1.22 repeatedly. First use this lemma to obtain V open such that its closure is compact and contained in U with $V \supseteq H$. Thus $H \subseteq V \subseteq \overline{V} \subseteq U, \overline{V}$ compact.

Let $D \equiv \{r_n\}_{n=1}^{\infty}$ be the rational numbers in $(0,1)$. Using Lemma 19.1.22, let V_{r_1} be an open set such that

$$H \subseteq V_{r_1} \subseteq \overline{V}_{r_1} \subseteq V, \ \overline{V}_{r_1} \text{ is compact}$$

Suppose $V_{r_1}, \cdots, V_{r_k}$ have been chosen and list the rational numbers $r_1, \cdots, r_k$ in order,

$$r_{l_1} < r_{l_2} < \cdots < r_{l_k} \text{ for } \{l_1, \cdots, l_k\} = \{1, \cdots, k\}.$$

If $r_{k+1} > r_{l_k}$ then letting $p = r_{l_k}$, let $V_{r_{k+1}}$ satisfy

$$\overline{V}_p \subseteq V_{r_{k+1}} \subseteq \overline{V}_{r_{k+1}} \subseteq V, \ \overline{V}_{r_{k+1}} \text{ compact}$$

If $r_{k+1} \in (r_{l_i}, r_{l_{i+1}})$, let $p = r_{l_i}$ and let $q = r_{l_{i+1}}$. Then let $V_{r_{k+1}}$ satisfy

$$\overline{V}_p \subseteq V_{r_{k+1}} \subseteq \overline{V}_{r_{k+1}} \subseteq V_q, \quad \overline{V}_{r_{k+1}} \text{ compact}$$

If $r_{k+1} < r_{l_1}$, let $p = r_{l_1}$ and let $V_{r_{k+1}}$ satisfy

$$H \subseteq V_{r_{k+1}} \subseteq \overline{V}_{r_{k+1}} \subseteq V_p, \ \overline{V}_{r_{k+1}} \text{ compact}$$

Thus there exist open sets V_r for each $r \in \mathbb{Q} \cap (0,1)$ with the property that if $r < s$,

$$H \subseteq V_r \subseteq \overline{V}_r \subseteq V_s \subseteq \overline{V}_s \subseteq V.$$

Now for $D \equiv \mathbb{Q} \cap (0,1)$, in the following, t will be in D

$$f(x) \equiv \min(\inf\{t \in D : x \in V_t\}, 1), \ f(x) \equiv 1 \text{ if } x \notin \bigcup_{t \in D} V_t.$$

I claim f is continuous. $f(x) < a$, means there must be some $t \in D$ with $t < a$ and $x \in V_t$. Thus

$$f^{-1}([0,a)) = \cup\{V_t : t < a, t \in D\},$$

an open set.

Next consider $x \in f^{-1}([0,a])$ so $f(x) \leq a$. If $t > a$, then $x \in V_t$ from the definition of $f(x)$. Thus

$$f^{-1}([0,a]) \subseteq \cap\{V_t : t > a\} = \cap\{\overline{V}_t : t > a\}$$

which is a closed set. If $x \in \cap\{V_t : t > a\}$, then $f(x) \leq a$ from the definition and so equality holds and $f^{-1}([0,a])$ is closed. This is also true if $a = 1$. In this case, $f^{-1}([0,1]) = X$. Therefore, for $a \geq 0$,

$$f^{-1}((a,1]) \cup f^{-1}([0,a]) = X$$

and so $f^{-1}((a,1])$ is an open set. It follows that f is continuous because $f^{-1}(a,b) = f^{-1}([0,b)) \cap f^{-1}((a,1])$ the intersection of two open sets. Since this is so for every interval, it follows that the inverse image of any open set is open and so f is continuous. Clearly $f(x) = 0$ on H. If $x \in V^C$, then $x \notin V_t$ for any $t \in D$ so $f(x) = 1$ on V^C. Let $g(x) = 1 - f(x)$. ■

In any metric space there is a much easier proof of the conclusion of Urysohn's lemma which applies. The following is Lemma 3.12.1 listed here for convenience.

Lemma 19.1.24 *Let S be a nonempty subset of a metric space, (X,d). Define*

$$f(x) \equiv \operatorname{dist}(x,S) \equiv \inf\{d(x,y) : y \in S\}.$$

Then f is continuous.

In a metric space it is all much easier.

Theorem 19.1.25 *Let (X,τ) be a locally compact metric space in which the closures of balls are compact and let $H \subseteq U$ where H is compact and U is open. Then there exists $g : X \to [0,1]$ such that g is continuous, $g(x) = 1$ on H and $g(x) = 0$ if $x \notin V$ for some open set V whose closure is compact.*

Proof: Let $\delta > 0$ be such that for all $h \in H$, $\operatorname{dist}(h,U^C) > \delta$. This exists because $h \to \operatorname{dist}(h,U^C)$ is continuous and so achieves its minimum on H which must be positive because U^C is closed. Now consider the balls $B(h,\delta)$. These cover the compact set H and so there are finitely many which do so. $B(h_1,\delta),\cdots,B(h_m,\delta)$ where the closure of each of these is compact. Also $\overline{B(h_j,\delta)} \subseteq U$. Because if $x \in \overline{B(h_j,\delta)}$, then $d(x,h_j) \leq \delta < \operatorname{dist}(h,U^C)$. Let $V = \cup_{j=1}^m B(h_j,\delta)$. Thus $\overline{V} = \cup_{j=1}^m \overline{B(h_j,\delta)}$ because there are only finitely many sets. Also $\overline{V}$ is compact because it is a finite union of compact sets. Now define

$$g(x) \equiv \frac{\operatorname{dist}(x,V^C)}{\operatorname{dist}(x,H) + \operatorname{dist}(x,V^C)}$$

This is continuous, equals 1 on H and equals 0 off V because the denominator is always positive since both H, V^C are closed. ■

A useful construction when dealing with locally compact Hausdorff spaces is the notion of the one point compactification of the space.

Definition 19.1.26 *Suppose (X,τ) is a locally compact Hausdorff space. Then let $\widetilde{X} \equiv X \cup \{\infty\}$ where ∞ is just the name of some point which is not in X which is called the point at infinity. A basis for the topology $\widetilde{\tau}$ for $\widetilde{X}$ is*

$$\tau \cup \left\{K^C \text{ where } K \text{ is a compact subset of } X\right\}.$$

The complement is taken with respect to $\widetilde{X}$ and so the open sets, K^C are basic open sets which contain ∞.

The reason this is called a compactification is contained in the next lemma.

Lemma 19.1.27 *If (X,τ) is a locally compact Hausdorff space, then $\left(\widetilde{X},\widetilde{\tau}\right)$ is a compact Hausdorff space. Also if U is an open set of $\widetilde{\tau}$, then $U \setminus \{\infty\}$ is an open set of τ.*

Proof: Since (X,τ) is a locally compact Hausdorff space, it follows $\left(\widetilde{X},\widetilde{\tau}\right)$ is a Hausdorff topological space. The only case which needs checking is the one of $p \in X$ and ∞. Since (X,τ) is locally compact, there exists an open set of τ, U having compact closure which contains p. Then $p \in U$ and $\infty \in \overline{U}^C$ and these are disjoint open sets containing the points, p and ∞ respectively. Now let $\mathscr{C}$ be an open cover of $\widetilde{X}$ with sets from $\widetilde{\tau}$. Then ∞ must be in some set, U_∞ from $\mathscr{C}$, which must contain a set of the form K^C where K is a compact subset of X. Then there exist sets from $\mathscr{C}$, $U_1,\cdots,U_r$ which cover K. Therefore, a finite subcover of $\widetilde{X}$ is $U_1,\cdots,U_r,U_\infty$.

To see the last claim, suppose U contains ∞ since otherwise there is nothing to show. Notice that if C is a compact set, then $X \setminus C$ is an open set. Therefore, if $x \in U \setminus \{\infty\}$, and if $\widetilde{X} \setminus C$ is a basic open set contained in U containing ∞, then if x is in this basic open set of $\widetilde{X}$, it is also in the open set $X \setminus C \subseteq U \setminus \{\infty\}$. If x is not in any basic open set of the form $\widetilde{X} \setminus C$ then x is contained in an open set of τ which is contained in $U \setminus \{\infty\}$. Thus $U \setminus \{\infty\}$ is indeed open in τ. ■

Lemma 19.1.28 *Let (X,τ) be a topological space and let $\mathscr{B}$ be a basis for τ. Then K is compact if and only if every open cover of basic open sets admits a finite subcover.*

Proof: Suppose first that X is compact. Then if $\mathscr{C}$ is an open cover consisting of basic open sets, it follows it admits a finite subcover because these are open sets in $\mathscr{C}$.

Next suppose that every basic open cover admits a finite subcover and let $\mathscr{C}$ be an open cover of X. Then define $\widetilde{\mathscr{C}}$ to be the collection of basic open sets which are contained in some set of $\mathscr{C}$. It follows $\widetilde{\mathscr{C}}$ is a basic open cover of X and so it admits a finite subcover, $\left\{U_1,\cdots,U_p\right\}$. Now each U_i is contained in an open set of $\mathscr{C}$. Let O_i be a set of $\mathscr{C}$ which contains U_i. Then $\left\{O_1,\cdots,O_p\right\}$ is an open cover of X. ■

Actually, there is a profound generalization of this lemma.

19.2 The Alexander Sub-basis Theorem

The Hausdorff maximal theorem is one of several convenient versions of the axiom of choice. For a discussion of this, see the appendix on the subject. There is this one, the well ordering principal, and Zorn's lemma. They are all equivalent to the axiom of choice and which one you use is a matter of taste.

Theorem 19.2.1 *(Hausdorff maximal principle) Let $\mathscr{F}$ be a nonempty partially ordered set. Then there exists a maximal chain.*

The main tool in the study of products of compact topological spaces is the Alexander subbasis theorem which is presented next. Recall a set is compact if every basic open cover admits a finite subcover, Lemma 19.1.28. This was pretty easy to prove. However, there is a much smaller set of open sets called a subbasis which has this property. The proof of this result is much harder.

Definition 19.2.2 *$\mathscr{S} \subseteq \tau$ is called a subbasis for the topology τ if the set $\mathscr{B}$ of finite intersections of sets of $\mathscr{S}$ is a basis for the topology τ.*

Theorem 19.2.3 *Let (X,τ) be a topological space and let $\mathscr{S} \subseteq \tau$ be a subbasis for τ. Then if $H \subseteq X$, H is compact if and only if every open cover of H consisting entirely of sets of $\mathscr{S}$ admits a finite subcover.*

Proof: The only if part is obvious because the subasic sets are themselves open.

If every basic open cover admits a finite subcover then the set in question is compact. Suppose then that H is a subset of X having the property that subbasic open covers admit finite subcovers. Is H compact? Assume this is not so. Then what was just observed about basic covers implies there exists a basic open cover of H, $\mathscr{O}$, which admits no finite subcover. Let $\mathscr{F}$ be defined as

$$\{\mathscr{O} : \mathscr{O} \text{ is a basic open cover of } H \text{ which admits no finite subcover}\}.$$

The assumption is that $\mathscr{F}$ is nonempty. Partially order $\mathscr{F}$ by set inclusion and use the Hausdorff maximal principle to obtain a maximal chain, $\mathscr{C}$, of such open covers and let $\mathscr{D} = \cup\mathscr{C}$. If $\mathscr{D}$ admits a finite subcover, then since $\mathscr{C}$ is a chain and the finite subcover has only finitely many sets, some element of $\mathscr{C}$ would also admit a finite subcover, contrary to the definition of $\mathscr{F}$. Therefore, $\mathscr{D}$ admits no finite subcover. If $\mathscr{D}'$ properly contains $\mathscr{D}$ and $\mathscr{D}'$ is a basic open cover of H, then $\mathscr{D}'$ has a finite subcover of H since otherwise, $\mathscr{C}$ would fail to be a maximal chain, being properly contained in $\mathscr{C} \cup \{\mathscr{D}'\}$. Every set of $\mathscr{D}$ is of the form

$$U = \cap_{i=1}^{m} B_i,\ B_i \in \mathscr{S}$$

because they are all basic open sets. If it is the case that for all $U \in \mathscr{D}$ one of the B_i is found in $\mathscr{D}$, then replace each such U with the subbasic set from $\mathscr{D}$ containing it. But then this would be a subbasic open cover of H which by assumption would admit a finite subcover contrary to the properties of $\mathscr{D}$. Therefore, one of the sets of $\mathscr{D}$, denoted by U, has the property that $U = \cap_{i=1}^{m} B_i$, $B_i \in \mathscr{S}$ and no B_i is in $\mathscr{D}$. Thus $\mathscr{D} \cup \{B_i\}$ admits a finite subcover, for each of the above B_i because it is strictly larger than $\mathscr{D}$. Let this finite subcover corresponding to B_i be denoted by $V_1^i, \cdots, V_{m_i}^i, B_i$. Consider $\{U, V_j^i,\ j = 1, \cdots, m_i,\ i = 1, \cdots, m\}$. If $p \in H \setminus \cup\{V_j^i\}$, then $p \in B_i$ for each i and so $p \in U$. This is therefore a finite subcover of $\mathscr{D}$ contradicting the properties of $\mathscr{D}$. Therefore, $\mathscr{F}$ must be empty. ■

19.3 The Product Topology and Compactness

Now here is the definition of the product topological space.

Definition 19.3.1 *Let (X_i,τ_i) for $i\in I$ be a topological space. By the axiom of choice, $\prod_{i\in I}X_i\neq\emptyset$. Then by definition, a basis for a topology on $\prod_{i\in I}X_i$ consists of sets of the form $\prod_{i\in I}A_i$ where $A_i=X_i$ except for* ***finitely*** *many i and for these, $A_i\in\tau_i$. A sub-basis for this topology consists of sets of the form $\prod_{i\in I}A_i$ where $A_i=X_i$ for all but a single i and for this one, $A_i\in\tau_i$. This product topology is denoted by $\prod\tau_i$. The resulting topological space is $(\prod_{i\in I}X_i,\prod\tau_i)$. "Subbasic" sets will be those which are in the sub-basis.*

It is important that the basic open sets have the i^{th} entries not all of X_i in only **finitely many** i. If you don't insist on this, you will be dealing with something called the "box topology" and the following major theorem will not be true. You might go through the proof and see that this is the case. By the Alexander subbasis theorem, compactness is equivalent to saying that every open cover of subbasic sets admits a finite subcover.

Theorem 19.3.2 *(Tychanoff) If (X_i,τ_i) is compact, then so is $(\prod_{i\in I}X_i,\prod\tau_i)$.*

Proof: By the Alexander subbasis theorem, the theorem will be proved if every subbasic open cover admits a finite subcover. Therefore, let $\mathscr{O}$ be a subbasic open cover of $\prod_{i\in I}X_i$. Let

$$\mathscr{O}_j=\{Q\in\mathscr{O}:Q=P_j(A)\text{ for some }A\in\tau_j\}.$$

Thus $\mathscr{O}_j$ consists of those sets of $\mathscr{O}$ which have a possibly proper subset of X_i only in the slot $i=j$. Let

$$\pi_j\mathscr{O}_j=\{A:P_j(A)\in\mathscr{O}_j\}.$$

Thus $\pi_j\mathscr{O}_j$ picks out those proper open subsets of X_j which occur in $\mathscr{O}_j$.

If no $\pi_j\mathscr{O}_j$ covers X_j, then by the axiom of choice, there exists $f\in\prod_{i\in I}X_i\setminus\cup\pi_i\mathscr{O}_i$. Therefore, $f(j)\notin\cup\pi_j\mathscr{O}_j$ for each $j\in I$. Now f is a point of $\prod_{i\in I}X_i$ and so $f\in P_k(A)\in\mathscr{O}$ for some k. However, this is a contradiction as it was shown that $f(k)$ is not an element of A. (A is one of the sets whose union makes up $\cup\pi_k\mathscr{O}_k$.) This contradiction shows that for some j, $\pi_j\mathscr{O}_j$ covers X_j. Thus $X_j=\cup\pi_j\mathscr{O}_j$ and so by compactness of X_j, there exist $A_1,\cdots,A_m$, sets in τ_j such that $X_j\subseteq\cup_{i=1}^m A_i$ and $P_j(A_i)\in\mathscr{O}$. Therefore, $\{P_j(A_i)\}_{i=1}^m$ covers $\prod_{i\in I}X_i$. By the Alexander subbasis theorem this proves $\prod_{i\in I}X_i$ is compact. ■

19.4 Stone Weierstrass Theorem

This theorem was presented earlier in the context of a real algebra of functions on a compact set. Here this is extended to the case where the functions are defined on a locally compact Hausdorff space and also extended to the case where the functions have values in $\mathbb{C}$.

19.4.1 The Case of Locally Compact Sets

Definition 19.4.1 *Let (X,τ) be a locally compact Hausdorff space. $C_0(X)$ denotes the space of real or complex valued continuous functions defined on X with the property that if $f\in C_0(X)$, then for each $\varepsilon>0$ there exists a compact set K such that $|f(x)|<\varepsilon$ for all $x\notin K$. Define $||f||_\infty=\sup\{|f(x)|:x\in X\}$.*

This norm is well defined because $|f(x)|<1$ for x not in some compact set K and $|f(x)|$ achieves its maximum on K.

Lemma 19.4.2 *For (X,τ) a locally compact Hausdorff space with the above norm, $C_0(X)$ is a complete space.*

Proof: Let $\left(\widetilde{X},\widetilde{\tau}\right)$ be the one point compactification described in Lemma 19.1.27.

$$D \equiv \left\{ f \in C\left(\widetilde{X}\right) : f(\infty) = 0 \right\}.$$

Then D is a closed subspace of $C\left(\widetilde{X}\right)$. For $f \in C_0(X)$,

$$\widetilde{f}(x) \equiv \begin{cases} f(x) \text{ if } x \in X \\ 0 \text{ if } x = \infty \end{cases}$$

and let $\theta : C_0(X) \to D$ be given by $\theta f = \widetilde{f}$. Then θ is one to one and onto and also satisfies $\|f\|_\infty = \|\theta f\|_\infty$. Now D is complete because it is a closed subspace of a complete space and so $C_0(X)$ with $\|\cdot\|_\infty$ is also complete. ■

The above refers to functions which have values in $\mathbb{C}$ but the same proof works for functions which have values in any complete normed linear space.

In the case where the functions in $C_0(X)$ all have real values, I will denote the resulting space by $C_0(X;\mathbb{R})$ with similar meanings in other cases.

With this lemma, the generalization of the Stone Weierstrass theorem to locally compact sets is as follows.

Theorem 19.4.3 *Let $\mathscr{A}$ be an algebra of functions in $C_0(X;\mathbb{R})$ where (X,τ) is a locally compact Hausdorff space which separates the points and annihilates no point. Then $\mathscr{A}$ is dense in $C_0(X;\mathbb{R})$.*

Proof: Let $\left(\widetilde{X},\widetilde{\tau}\right)$ be the one point compactification as described in Lemma 19.1.27. Let $\widetilde{\mathscr{A}}$ denote all finite linear combinations, $\left\{\sum_{i=1}^n c_i \widetilde{f}_i + c_0 : f \in \mathscr{A},\ c_i \in \mathbb{R}\right\}$ where for $f \in C_0(X;\mathbb{R})$,

$$\widetilde{f}(x) \equiv \begin{cases} f(x) \text{ if } x \in X \\ 0 \text{ if } x = \infty \end{cases}.$$

Then $\widetilde{\mathscr{A}}$ is obviously an algebra of functions in $C\left(\widetilde{X};\mathbb{R}\right)$. It separates points because this is true of $\mathscr{A}$. Similarly, it annihilates no point because of the inclusion of c_0 an arbitrary element of $\mathbb{R}$ in the definition above. Therefore from Theorem 5.9.5, $\widetilde{\mathscr{A}}$ is dense in $C\left(\widetilde{X};\mathbb{R}\right)$. Letting $f \in C_0(X;\mathbb{R})$, it follows $\widetilde{f} \in C\left(\widetilde{X};\mathbb{R}\right)$ so there exists a sequence $\{h_n\} \subseteq \widetilde{\mathscr{A}}$ such that h_n converges uniformly to $\widetilde{f}$. Now h_n is of the form $\sum_{i=1}^n c_i^n \widetilde{f}_i^n + c_0^n$ and since $\widetilde{f}(\infty) = 0$, you can take each $c_0^n = 0$ and so this has shown the existence of a sequence of functions in $\mathscr{A}$ such that it converges uniformly to f. ■

19.4.2 The Case of Complex Valued Functions

What about the general case where $C_0(X)$ consists of complex valued functions and the field of scalars is $\mathbb{C}$ rather than $\mathbb{R}$? The following is the version of the Stone Weierstrass theorem which applies to this case. You have to assume that for $f \in \mathscr{A}$ it follows $\bar{f} \in \mathscr{A}$.

Lemma 19.4.4 *Let z be a complex number. Then* $\mathrm{Re}(z) = \mathrm{Im}(i\bar{z})$, $\mathrm{Im}(z) = \mathrm{Re}(i\bar{z})$.

Proof: The following computation comes from the definition of real and imaginary parts.

$$\begin{aligned}\mathrm{Re}(z) &= \frac{z+\bar{z}}{2} = \frac{iz+i\bar{z}}{2i} = \frac{i\bar{z}-\overline{(i\bar{z})}}{2i} = \mathrm{Im}(i\bar{z}) \\ \mathrm{Im}(z) &= \frac{z-\bar{z}}{2i} = \frac{i\bar{z}-iz}{2} = \frac{i\bar{z}+\overline{(i\bar{z})}}{2} = \mathrm{Re}(i\bar{z}) \ \blacksquare\end{aligned}$$

Theorem 19.4.5 *Suppose $\mathscr{A}$ is an algebra of functions in $C_0(X)$ for X a locally compact Hausdorff space which separates the points of X and annihilates no point of X, and has the property that if $f \in \mathscr{A}$, then $\bar{f} \in \mathscr{A}$. Then $\mathscr{A}$ is dense in $C_0(X)$.*

Proof: Let $\mathrm{Re}\,\mathscr{A} \equiv \{\mathrm{Re}\, f : f \in \mathscr{A}\}$, $\mathrm{Im}\,\mathscr{A} \equiv \{\mathrm{Im}\, f : f \in \mathscr{A}\}$.

Claim 1: $\mathrm{Re}\,\mathscr{A} = \mathrm{Im}\,\mathscr{A}$

Proof of claim: A typical element of $\mathrm{Re}\,\mathscr{A}$ is $\mathrm{Re}\, f$ where $f \in \mathscr{A}$, then from Lemma 19.4.4, $\mathrm{Re}(f) = \mathrm{Im}\left(i\bar{f}\right) \in \mathrm{Im}\,\mathscr{A}$. Thus $\mathrm{Re}\,\mathscr{A} \subseteq \mathrm{Im}\,\mathscr{A}$. By assumption, $i\bar{f} \in \mathscr{A}$. The other direction works the same. Just use the other formula in Lemma 19.4.4.

Claim 2: Both $\mathrm{Re}\,\mathscr{A}$ and $\mathrm{Im}\,\mathscr{A}$ are real algebras.

Proof of claim: It is obvious these are both real vector spaces. Since these are equal, it suffices to consider $\mathrm{Re}\,\mathscr{A}$. It remains to show that $\mathrm{Re}\,\mathscr{A}$ is closed with respect to products.

$$\frac{f+\bar{f}}{2}\frac{g+\bar{g}}{2} = \frac{1}{4}\left[fg+f\bar{g}+\bar{f}g+\overline{fg}\right] = \frac{1}{4}\left[2\mathrm{Re}(fg)+2\mathrm{Re}\left(\bar{f}g\right)\right]$$

Now by assumption, $fg \in \mathscr{A}$ and so $\mathrm{Re}(fg) \in \mathrm{Re}\,\mathscr{A}$. Also $\mathrm{Re}\left(\bar{f}g\right) \in \mathrm{Re}\,\mathscr{A}$ because both $\bar{f}, g$ are in $\mathscr{A}$ and it is an algebra. Thus, the above is in $\mathrm{Re}\,\mathscr{A}$ because, as noted, this is a real vector space.

Claim 3: $\mathscr{A} = \mathrm{Re}\,\mathscr{A} + i\,\mathrm{Im}\,\mathscr{A}$

Proof of claim: If $f \in \mathscr{A}$, then $f = \frac{f+\bar{f}}{2} + i\frac{f-\bar{f}}{2i} \in \mathrm{Re}\,\mathscr{A} + i\,\mathrm{Im}\,\mathscr{A}$ so $\mathscr{A} \subseteq \mathrm{Re}\,\mathscr{A} + i\,\mathrm{Im}\,\mathscr{A}$. Now a generic element of $\mathrm{Re}\,\mathscr{A} + i\,\mathrm{Im}\,\mathscr{A}$ is $\mathrm{Re}(f) + i\,\mathrm{Im}(g)$ for $f, g \in \mathscr{A}$.

$$\mathrm{Re}(f) + i\,\mathrm{Im}(g) \equiv \frac{f+\bar{f}}{2} + i\left(\frac{g-\bar{g}}{2i}\right) = \frac{f+g}{2} + \frac{\bar{f}-\bar{g}}{2} \in \mathscr{A}$$

because $\mathscr{A}$ is closed with respect to conjugates. Thus $\mathrm{Re}\,\mathscr{A} + i\,\mathrm{Im}\,\mathscr{A} \subseteq \mathscr{A}$.

Both $\mathrm{Re}\,\mathscr{A}$ and $\mathrm{Im}\,\mathscr{A}$ must separate the points. Here is why: If $x_1 \neq x_2$, then there exists $f \in \mathscr{A}$ such that $f(x_1) \neq f(x_2)$. If $\mathrm{Im}\, f(x_1) \neq \mathrm{Im}\, f(x_2)$, this shows there is a function in $\mathrm{Im}\,\mathscr{A}$, $\mathrm{Im}\, f$ which separates these two points. If $\mathrm{Im}\, f$ fails to separate the two points, then $\mathrm{Re}\, f$ must separate the points and so, by Lemma 19.4.4,

$$\mathrm{Re}\, f(x_1) = \mathrm{Im}\left(i\bar{f}(x_1)\right) \neq \mathrm{Re}\, f(x_2) = \mathrm{Im}\left(i\bar{f}(x_2)\right)$$

Thus $\mathrm{Im}\,\mathscr{A}$ separages the points. Similarly $\mathrm{Re}\,\mathscr{A}$ separates the points using a similar argument or because it is equal to $\mathrm{Im}\,\mathscr{A}$.

Neither $\mathrm{Re}\,\mathscr{A}$ nor $\mathrm{Im}\,\mathscr{A}$ annihilate any point. This is easy to see because if x is a point, there exists $f \in \mathscr{A}$ such that $f(x) \neq 0$. Thus either $\mathrm{Re}\, f(x) \neq 0$ or $\mathrm{Im}\, f(x) \neq 0$. If $\mathrm{Im}\, f(x) \neq 0$, this shows this point is not annihilated by $\mathrm{Im}\,\mathscr{A}$. Since they are equal, $\mathrm{Re}\,\mathscr{A}$ does not annihilate this point either.

It follows from Theorem 19.4.3 that $\mathrm{Re}\,\mathscr{A}$ and $\mathrm{Im}\,\mathscr{A}$ are dense in the real valued functions of $C_0(X)$. Let $f \in C_0(X)$. Then there exists $\{h_n\} \subseteq \mathrm{Re}\,\mathscr{A}$ and $\{g_n\} \subseteq \mathrm{Im}\,\mathscr{A}$ such that $h_n \to \mathrm{Re}\, f$ uniformly and $g_n \to \mathrm{Im}\, f$ uniformly. Therefore, $h_n + ig_n \in \mathscr{A}$ and it converges to f uniformly. $\blacksquare$

19.5 Partitions of Unity

As before, the idea of a partition of unity if of fundamental significance. It will be used to construct measures.

Definition 19.5.1 *Define* $\mathrm{spt}(f)$ *(support of* f*) to be the closure of the set* $\{x : f(x) \neq 0\}$. *If* V *is an open set,* $C_c(V)$ *will be the set of continuous functions* f*, defined on* Ω *having* $\mathrm{spt}(f) \subseteq V$. *Thus in Theorem 19.1.23,* $f \in C_c(V)$.

Definition 19.5.2 *If* K *is a compact subset of an open set* V*, then* $K \prec \phi \prec V$ *if*

$$\phi \in C_c(V),\ \phi(K) = \{1\},\ \phi(\Omega) \subseteq [0,1],$$

where Ω *denotes the whole topological space considered. Also for* $\phi \in C_c(\Omega)$, $K \prec \phi$ *if*

$$\phi(\Omega) \subseteq [0,1] \text{ and } \phi(K) = 1.$$

and $\phi \prec V$ *if* $\phi \in C_c(V)$ *and*

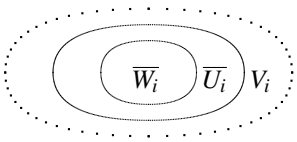
$$\phi(\Omega) \subseteq [0,1] \text{ and } \mathrm{spt}(\phi) \subseteq V.$$

Theorem 19.5.3 *(Partition of unity) Let* K *be a compact subset of a locally compact Hausdorff topological space satisfying Theorem 19.1.23 and suppose*

$$K \subseteq V = \cup_{i=1}^n V_i,\ V_i \text{ open.}$$

Then there exist $\psi_i \prec V_i$ *with* $\sum_{i=1}^n \psi_i(x) = 1$ *for all* $x \in K$.

Proof: The proof is just like the one in Theorem 3.12.5 on Page 92. Let $K_1 = K \setminus \cup_{i=2}^n V_i$. Thus K_1 is compact and $K_1 \subseteq V_1$. Let $K_1 \subseteq W_1 \subseteq \overline{W}_1 \subseteq V_1$ with $\overline{W}_1$ compact. To obtain W_1, use Theorem 19.1.23 to get f such that $K_1 \prec f \prec V_1$ and let $W_1 \equiv \{x : f(x) \neq 0\}$. Thus $W_1, V_2, \cdots V_n$ covers K and $\overline{W}_1 \subseteq V_1$. Let $K_2 = K \setminus (\cup_{i=3}^n V_i \cup W_1)$. Then K_2 is compact and $K_2 \subseteq V_2$. Let $K_2 \subseteq W_2 \subseteq \overline{W}_2 \subseteq V_2$, $\overline{W}_2$ compact. Continue this way finally obtaining $W_1, \cdots, W_n$, $K \subseteq W_1 \cup \cdots \cup W_n$, and $\overline{W}_i \subseteq V_i$ $\overline{W}_i$ compact. Now let $\overline{W}_i \subseteq U_i \subseteq \overline{U}_i \subseteq V_i$, $\overline{U}_i$ compact.



By Theorem 19.1.23, let $\overline{U}_i \prec \phi_i \prec V_i$, $\cup_{i=1}^n \overline{W}_i \prec \gamma \prec \cup_{i=1}^n U_i$. Define

$$\psi_i(x) = \begin{cases} \gamma(x)\phi_i(x)/\sum_{j=1}^n \phi_j(x) \text{ if } \sum_{j=1}^n \phi_j(x) \neq 0, \\ 0 \text{ if } \sum_{j=1}^n \phi_j(x) = 0. \end{cases}$$

If x is such that $\sum_{j=1}^n \phi_j(x) = 0$, then $x \notin \cup_{i=1}^n \overline{U}_i$. Consequently $\gamma(y) = 0$ for all y near x and so $\psi_i(y) = 0$ for all y near x. Hence ψ_i is continuous at such x. If $\sum_{j=1}^n \phi_j(x) \neq 0$, this situation persists near x and so ψ_i is continuous at such points. Therefore ψ_i is continuous. If $x \in K$, then $\gamma(x) = 1$ and so $\sum_{j=1}^n \psi_j(x) = 1$. Clearly $0 \leq \psi_i(x) \leq 1$ and $\mathrm{spt}(\psi_j) \subseteq V_j$. ■

The following corollary won't be needed immediately but is quite useful.

Corollary 19.5.4 *In the context of the above theorem, if* H *is a compact subset of* V_i*, there exists a partition of unity such that* $\psi_i(x) = 1$ *for all* $x \in H$ *in addition to the conclusion of Theorem 19.5.3.*

Proof: Keep V_i the same but replace V_j with $\widetilde{V}_j \equiv V_j \setminus H$. Now in the proof above, applied to this modified collection of open sets, if $j \neq i$, $\phi_j(x) = 0$ whenever $x \in H$. Therefore, $\psi_i(x) = 1$ on H. ■

19.6 Measures on Hausdorff Spaces

In the case of a Hausdorff topological space, the following lemma gives conditions under which the σ algebra of μ measurable sets for an outer measure μ contains the Borel sets. In words, it assumes the outer measure is inner regular on open sets and outer regular on all sets. Also it assumes you can approximate the measure of an open set with a compact set and the measure of a compact set with an open set. Recall that the Borel sets are those sets in the smallest σ algebra that contains the open sets. The big result is the Riesz representation theorem for positive linear functionals and the following lemma is the technical part of the proof of this big theorem in addition to being interesting for its own sake. It holds in a Hausdorff space, not just one which is locally compact.

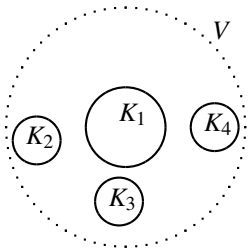
Lemma 19.6.1 *Let Ω be a Hausdorff space and suppose μ is an outer measure satisfying μ is finite on compact sets and the following conditions,*

1. $\mu(E) = \inf\{\mu(V), V \supseteq E, V \text{ open}\}$ *for all E. (Outer regularity.)*
2. *For every open set V, $\mu(V) = \sup\{\mu(K) : K \subseteq V, K \text{ compact}\}$ (Inner regularity on open sets.)*
3. *If A, B are compact disjoint sets, then $\mu(A \cup B) = \mu(A) + \mu(B)$.*

Then the following hold.

1. *If $\varepsilon > 0$ and if K is compact, there exists V open such that $V \supseteq K$ and*
$$\mu(V \setminus K) < \varepsilon$$
2. *If $\varepsilon > 0$ and if V is open with $\mu(V) < \infty$, there exists a compact subset K of V such that*
$$\mu(V \setminus K) < \varepsilon$$
3. *The μ measurable sets $\mathscr{S}$ defined as*
$$\mathscr{S} \equiv \{E \subseteq \Omega : \mu(S) = \mu(S \setminus E) + \mu(S \cap E) \text{ for all } S\}$$
contains the Borel sets and also μ is inner regular on every open set,
$$\mu(V) = \sup\{\mu(K) : K \subseteq V, K \text{ compact}\}$$
and for every $E \in \mathscr{S}$ with $\mu(E) < \infty$,
$$\mu(E) = \sup\{\mu(K) : K \subseteq E, K \text{ compact}\}$$

Proof: First we establish 1 and 2 and use them to establish the last assertion. Consider 2. Suppose it is not true. Then there exists an open set V having $\mu(V) < \infty$ but for all $K \subseteq V$, $\mu(V \setminus K) \geq \varepsilon$ for some $\varepsilon > 0$. By inner regularity on open sets, there exists $K_1 \subseteq V, K_1$ compact, such that $\mu(K_1) \geq \varepsilon/2$. Now by assumption, $\mu(V \setminus K_1) \geq \varepsilon$ and so by inner regularity on open sets again, there exists compact $K_2 \subseteq V \setminus K_1$ such that $\mu(K_2) \geq \varepsilon/2$. Continuing this way, there is a sequence of disjoint compact sets contained in V $\{K_i\}$ such that $\mu(K_i) \geq \varepsilon/2$.



Now this is an obvious contradiction because by 3,

$$\mu(V) \geq \mu\left(\cup_{i=1}^{n} K_i\right) = \sum_{i=1}^{n} \mu(K_i) \geq n\frac{\varepsilon}{2}$$

for each n, contradicting $\mu(V) < \infty$.

Next consider 1. By outer regularity, there exists an open set $W \supseteq K$ such that $\mu(W) < \mu(K) + 1$. By 2, there exists compact $K_1 \subseteq W \setminus K$ such that $\mu((W \setminus K) \setminus K_1) < \varepsilon$. Then consider $V \equiv W \setminus K_1$. This is an open set containing K and from what was just shown,

$$\mu((W \setminus K_1) \setminus K) = \mu((W \setminus K) \setminus K_1) < \varepsilon.$$

Now consider the last assertion.

Define $\mathscr{S}_1 = \{E \in \mathscr{P}(\Omega) : E \cap K \in \mathscr{S}\}$ for all compact K.

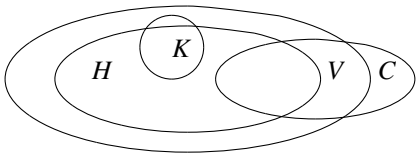
First it will be shown the compact sets are in $\mathscr{S}$. From this it will follow the closed sets are in $\mathscr{S}_1$. Then you show $\mathscr{S}_1 = \mathscr{S}$. Thus $\mathscr{S}_1 = \mathscr{S}$ is a σ algebra and so it contains the Borel sets since it contains the closed sets. Finally you show the inner regularity assertion.

Claim 1: Compact sets are in $\mathscr{S}$.

Proof of claim: Let V be an open set with $\mu(V) < \infty$. I will show that for C compact,

$$\mu(V) \geq \mu(V \setminus C) + \mu(V \cap C).$$

If $\mu(V) = \infty$ the above is obvious. The various sets are illustrated in the following diagram.



By 2, there exists a compact set $K \subseteq V \setminus C$ such that $\mu((V \setminus C) \setminus K) < \varepsilon$ and a compact set $H \subseteq V$ such that $\mu(V \setminus H) < \varepsilon$. Thus $\mu(V) \leq \mu(V \setminus H) + \mu(H) < \varepsilon + \mu(H)$. Then

$$\mu(V) \leq \mu(H) + \varepsilon \leq \mu(H \cap C) + \mu(H \setminus C) + \varepsilon$$

$$\leq \mu(V \cap C) + \mu(V \setminus C) + \varepsilon \leq \mu(H \cap C) + \mu(K) + 3\varepsilon$$

By 3,

$$= \mu(H \cap C) + \mu(K) + 3\varepsilon = \mu((H \cap C) \cup K) + 3\varepsilon \leq \mu(V) + 3\varepsilon.$$

Since ε is arbitrary, this shows that

$$\mu(V) = \mu(V \setminus C) + \mu(V \cap C). \tag{19.4}$$

Of course 19.4 is exactly what needs to be shown for arbitrary S in place of V. It suffices to consider only S having $\mu(S) < \infty$. If $S \subseteq \Omega$, with $\mu(S) < \infty$, let $V \supseteq S$, $\mu(S) + \varepsilon > \mu(V)$. Then from what was just shown, if C is compact,

$$\varepsilon + \mu(S) > \mu(V) = \mu(V \setminus C) + \mu(V \cap C) \geq \mu(S \setminus C) + \mu(S \cap C).$$

Since ε is arbitrary, this shows the compact sets are in $\mathcal{S}$. This proves the claim.

As discussed above, this verifies the closed sets are in $\mathcal{S}_1$ because if H is closed and C is compact, then compact $H \cap C \in \mathcal{S}$. If $\mathcal{S}_1$ is a σ algebra, this will show that $\mathcal{S}_1$ contains the Borel sets. Thus I first show $\mathcal{S}_1$ is a σ algebra.

To see that $\mathcal{S}_1$ is closed with respect to taking complements, let $E \in \mathcal{S}_1$ and K a compact set.

$$K = (E^C \cap K) \cup (E \cap K).$$

Then from the fact, just established, that the compact sets are in $\mathcal{S}$, $E^C \cap K = K \setminus (E \cap K) \in \mathcal{S}$. $\mathcal{S}_1$ is closed under countable unions because if K is a compact set and $E_n \in \mathcal{S}_1$,

$$K \cap \cup_{n=1}^{\infty} E_n = \cup_{n=1}^{\infty} K \cap E_n \in \mathcal{S}$$

because it is a countable union of sets of $\mathcal{S}$. Thus $\mathcal{S}_1$ is a σ algebra.

Therefore, if $E \in \mathcal{S}$ and K is a compact set, just shown to be in $\mathcal{S}$, it follows $K \cap E \in \mathcal{S}$ because $\mathcal{S}$ is a σ algebra which contains the compact sets and so $\mathcal{S}_1 \supseteq \mathcal{S}$. It remains to verify $\mathcal{S}_1 \subseteq \mathcal{S}$. Recall that

$$\mathcal{S}_1 \equiv \{E : E \cap K \in \mathcal{S} \text{ for all } K \text{ compact}\}$$

Let $E \in \mathcal{S}_1$ and let V be an open set with $\mu(V) < \infty$ and choose $K \subseteq V$ such that $\mu(V \setminus K) < \varepsilon$. Then since $E \in \mathcal{S}_1$, it follows $E \cap K, E^C \cap K \in \mathcal{S}$ and so

$$\begin{aligned} \mu(V) &\leq \mu(V \setminus E) + \mu(V \cap E) \leq \overbrace{\mu(K \setminus E) + \mu(K \cap E)}^{\text{The two sets are disjoint and in } \mathcal{S}} + 2\varepsilon \\ &= \mu(K) + 2\varepsilon \leq \mu(V) + 3\varepsilon \end{aligned}$$

Since ε is arbitrary, this shows $\mu(V) = \mu(V \setminus E) + \mu(V \cap E)$ which would show $E \in \mathcal{S}$ if V were an arbitrary set.

Now let $S \subseteq \Omega$ be such an arbitrary set. If $\mu(S) = \infty$, then $\mu(S) = \mu(S \cap E) + \mu(S \setminus E)$. If $\mu(S) < \infty$, let

$$V \supseteq S, \ \mu(S) + \varepsilon \geq \mu(V).$$

Then

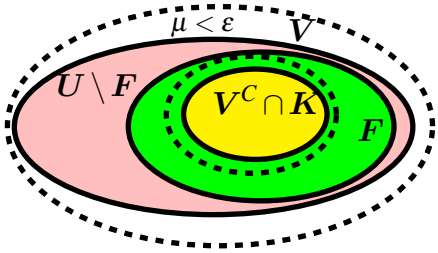
$$\mu(S) + \varepsilon \geq \mu(V) = \mu(V \setminus E) + \mu(V \cap E) \geq \mu(S \setminus E) + \mu(S \cap E).$$

Since ε is arbitrary, this shows that $E \in \mathcal{S}$ and so $\mathcal{S}_1 = \mathcal{S}$. Thus $\mathcal{S} \supseteq$ Borel sets as claimed.

From 2 μ is inner regular on all open sets. It remains to show that

$$\mu(F) = \sup\{\mu(K) : K \subseteq F\} \tag{19.5}$$

for all $F \in \mathcal{S}$ with $\mu(F) < \infty$. It might help to refer to the following crude picture to keep things straight. It also might not help. V is between the dotted lines.



From the picture as needed: Let $\mu(U \setminus F) < \varepsilon$ where U is open and let $K \subseteq U$ and $\mu(U \setminus K) < \varepsilon$, $\mu(V \setminus (U \setminus F)) < \varepsilon$ with V open and $V \supseteq U \setminus F = U \cap F^C$ so $V^C \subseteq U^C \cup F$. This is possible because all sets are in $\mathscr{S}$. Then $V^C \cap K \subseteq \left(U^C \cup F\right) \cap K = F \cap K \subseteq F$. Now $V^C \cap K$ is compact and

$$\begin{aligned} \mu\left(F \setminus \left(K \cap V^C\right)\right) &= \mu\left(F \cap \left(K^C \cup V\right)\right) = \mu(F \cap V) + \mu\left(F \cap K^C\right) \\ &\leq \mu(F \cap V) + \mu(U \setminus K) < \mu(F \cap V) + \varepsilon \end{aligned} \tag{19.6}$$

However,

$$\varepsilon > \mu(V \setminus (U \setminus F)) = \mu\left(V \cap \left(U \cap F^C\right)^C\right) = \mu\left(V \cap \left(U^C \cup F\right)\right) \geq \mu(V \cap F)$$

and so from 19.6, $\mu\left(F \setminus \left(K \cap V^C\right)\right) \leq 2\varepsilon$. Since $K \cap V^C$ is compact, this shows 19.5. ■

19.7 Measures and Positive Linear Functionals

This is on the Riesz representation theorem for positive linear functionals. It is a really marvelous result. It produces measures on locally compact Hausdorff spaces. Thus this doesn't help a lot in producing measures on infinite dimensional spaces but it works great on $\mathbb{R}^n$ or closed subsets of $\mathbb{R}^n$ and so forth.

Definition 19.7.1 *Let (Ω, τ) be a topological space. $L : C_c(\Omega) \to \mathbb{C}$ is called a positive linear functional if L is linear, $L(af_1 + bf_2) = aLf_1 + bLf_2$, and if $Lf \geq 0$ whenever $f \geq 0$.*

Theorem 19.7.2 *(Riesz representation theorem) Let (Ω, τ) be a locally compact Hausdorff space and let L be a positive linear functional on $C_c(\Omega)$. Then there exists a σ algebra $\mathscr{S}$ containing the Borel sets and a unique measure μ, defined on $\mathscr{S}$, such that*

$$\mu \text{ is complete,} \tag{19.7}$$

$$\mu(K) < \infty \text{ for all } K \text{ compact,} \tag{19.8}$$

$$\mu(F) = \sup\{\mu(K) : K \subseteq F,\ K \text{ compact}\},$$

for all F open and for all $F \in \mathscr{S}$ with $\mu(F) < \infty$,

$$\mu(F) = \inf\{\mu(V) : V \supseteq F,\ V \text{ open}\}$$

for all $F \in \mathscr{S}$, and

$$\int f d\mu = Lf \text{ for all } f \in C_c(\Omega). \tag{19.9}$$

The plan is to define an outer measure and then to show that it, together with the σ algebra of sets measurable in the sense of Caratheodory, satisfies the conclusions of the theorem. Always, K will be a compact set and V will be an open set.

Definition 19.7.3 *$\mu(V) \equiv \sup\{Lf : f \prec V\}$ for V open,*

$$\mu(\emptyset) = 0,\ \mu(E) \equiv \inf\{\mu(V) : V \supseteq E\}$$

for arbitrary sets E.

Lemma 19.7.4 *μ is a well-defined outer measure.*

Proof: First it is necessary to verify that μ is well defined because there are two descriptions of it on open sets. Suppose then that $\mu_1(V) \equiv \inf\{\mu(U) : U \supseteq V \text{ and } U \text{ is open}\}$. It is required to verify that $\mu_1(V) = \mu(V)$ where μ is given as $\sup\{Lf : f \prec V\}$. If $U \supseteq V$, then $\mu(U) \geq \mu(V)$ directly from the definition. Hence from the definition of μ_1, it follows $\mu_1(V) \geq \mu(V)$. On the other hand, $V \supseteq V$ and so $\mu_1(V) \leq \mu(V)$. This verifies μ is well defined.

It remains to show that μ is an outer measure. Let $V = \cup_{i=1}^{\infty} V_i$ and let $f \prec V$. Then $\text{spt}(f) \subseteq \cup_{i=1}^{n} V_i$ for some n. Let $\psi_i \prec V_i$, $\sum_{i=1}^{n} \psi_i = 1$ on $\text{spt}(f)$.

$$Lf = \sum_{i=1}^{n} L(f\psi_i) \leq \sum_{i=1}^{n} \mu(V_i) \leq \sum_{i=1}^{\infty} \mu(V_i).$$

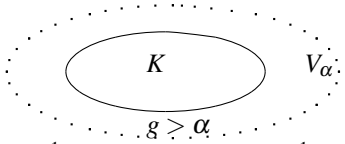
Hence $\mu(V) \leq \sum_{i=1}^{\infty} \mu(V_i)$ since $f \prec V$ is arbitrary. Now let $E = \cup_{i=1}^{\infty} E_i$. Is $\mu(E) \leq \sum_{i=1}^{\infty} \mu(E_i)$? Without loss of generality, it can be assumed $\mu(E_i) < \infty$ for each i since if not so, there is nothing to prove. Let $V_i \supseteq E_i$ with $\mu(E_i) + \varepsilon 2^{-i} > \mu(V_i)$.

$$\mu(E) \leq \mu(\cup_{i=1}^{\infty} V_i) \leq \sum_{i=1}^{\infty} \mu(V_i) \leq \varepsilon + \sum_{i=1}^{\infty} \mu(E_i).$$

Since ε was arbitrary, $\mu(E) \leq \sum_{i=1}^{\infty} \mu(E_i)$. It is clear from the definition that if $A \subseteq B$, then $\mu(A) \leq \mu(B)$. ■

Lemma 19.7.5 *Let K be compact, $g \geq 0$, $g \in C_c(\Omega)$, and $g = 1$ on K. Then $\mu(K) \leq Lg$. Also $\mu(K) < \infty$ whenever K is compact.*

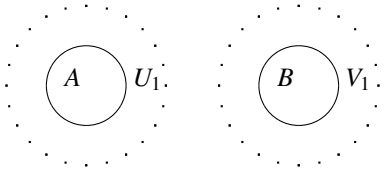
Proof: Let $\alpha \in (0,1)$ and $V_\alpha = \{x : g(x) > \alpha\}$ so $V_\alpha \supseteq K$ and let $h \prec V_\alpha$.



Then $h \leq 1$ on V_α while $g\alpha^{-1} \geq 1$ on V_α and so $g\alpha^{-1} \geq h$ which implies $L(g\alpha^{-1}) \geq Lh$ and that therefore, since L is linear, $Lg \geq \alpha Lh$. Since $h \prec V_\alpha$ is arbitrary, and $K \subseteq V_\alpha$, $Lg \geq \alpha\mu(V_\alpha) \geq \alpha\mu(K)$. Letting $\alpha \uparrow 1$ yields $Lg \geq \mu(K)$. This proves the first part of the lemma. The second assertion follows from this and Theorem 19.1.23. If K is given, let $K \prec g \prec \Omega$ and so from what was just shown, $\mu(K) \leq Lg < \infty$. ■

Lemma 19.7.6 *If A and B are disjoint compact subsets of Ω, then $\mu(A \cup B) = \mu(A) + \mu(B)$.*

Proof: By Theorem 19.1.23, there exists $h \in C_c(\Omega)$ such that $A \prec h \prec B^C$. Let $U_1 = h^{-1}((\frac{1}{2},1])$, $V_1 = h^{-1}([0,\frac{1}{2}))$. Then $A \subseteq U_1, B \subseteq V_1$ and $U_1 \cap V_1 = \emptyset$.



From Lemma 19.7.5 $\mu(A \cup B) < \infty$ and so there exists an open set, W such that

$$W \supseteq A \cup B,\ \mu(A \cup B) + \varepsilon > \mu(W).$$

Now let $U = U_1 \cap W$ and $V = V_1 \cap W$. Then

$$U \supseteq A,\ V \supseteq B,\ U \cap V = \emptyset, \text{and } \mu(A \cup B) + \varepsilon \geq \mu(W) \geq \mu(U \cup V).$$

Let $A \prec f \prec U$, $B \prec g \prec V$. Then by Lemma 19.7.5,

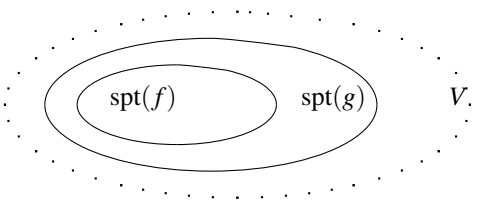
$$\mu(A \cup B) + \varepsilon \geq \mu(U \cup V) \geq L(f+g) = Lf + Lg \geq \mu(A) + \mu(B).$$

Since $\varepsilon > 0$ is arbitrary, this proves the lemma. ■

From Lemma 19.7.5 the following lemma is obtained.

Lemma 19.7.7 *Let $f \in C_c(\Omega)$, $f(\Omega) \subseteq [0,1]$. Then $\mu(\text{spt}(f)) \geq Lf$. Also, every open set, V satisfies $\mu(V) = \sup\{\mu(K) : K \subseteq V\}$.*

Proof: Let $V \supseteq \text{spt}(f)$ and let $\text{spt}(f) \prec g \prec V$. Then $Lf \leq Lg \leq \mu(V)$ because $f \leq g$. Since this holds for all $V \supseteq \text{spt}(f)$, $Lf \leq \mu(\text{spt}(f))$ by definition of μ. Finally, let V be open and let $l < \mu(V)$. Then from the definition of μ, there exists $f \prec V$ such that $L(f) > l$. Therefore, $l < \mu(\text{spt}(f)) \leq \mu(V)$ and so this shows the claim about inner regularity of the measure on an open set. ■



At this point, the conditions of Lemma 19.6.1 have been verified. Thus $\mathscr{S}$ contains the Borel sets and μ is inner regular on sets of $\mathscr{S}$ having finite measure.

It remains to show μ satisfies 19.9.

Lemma 19.7.8 *$\int f d\mu = Lf$ for all $f \in C_c(\Omega)$.*

Proof: Let $f \in C_c(\Omega)$, f real-valued, and suppose $f(\Omega) \subseteq [a,b]$. Choose $t_0 < a$ and let $t_0 < t_1 < \cdots < t_n = b$, $t_i - t_{i-1} < \varepsilon$. Let

$$E_i = f^{-1}((t_{i-1}, t_i]) \cap \text{spt}(f). \tag{19.10}$$

Note that $\cup_{i=1}^n E_i = \text{spt}(f)$ since $\Omega = \cup_{i=1}^n f^{-1}((t_{i-1}, t_i])$. Let $V_i \supseteq E_i, V_i$ is open and let V_i satisfy

$$f(x) < t_i + \varepsilon \text{ for all } x \in V_i,\ \mu(V_i \setminus E_i) < \varepsilon/n. \tag{19.11}$$

By Theorem 19.5.3 there exists $h_i \in C_c(\Omega)$ such that

$$h_i \prec V_i, \ \sum_{i=1}^{n} h_i(x) = 1 \text{ on } \operatorname{spt}(f).$$

Now note that for each i, $f(x)h_i(x) \leq h_i(x)(t_i+\varepsilon)$. (If $x \in V_i$, this follows from 19.11. If $x \notin V_i$ both sides equal 0.) Therefore,

$$\begin{aligned} Lf &= L(\sum_{i=1}^{n} fh_i) \leq L(\sum_{i=1}^{n} h_i(t_i+\varepsilon)) = \sum_{i=1}^{n}(t_i+\varepsilon)L(h_i) \\ &= \sum_{i=1}^{n}(|t_0|+t_i+\varepsilon)L(h_i) - |t_0|L\left(\overbrace{\sum_{i=1}^{n} h_i}^{=1 \text{ on } \operatorname{spt}(f)}\right). \end{aligned}$$

Now note that $|t_0|+t_i+\varepsilon \geq 0$ and so from the definition of μ and Lemma 19.7.5, this is no larger than

$$\begin{aligned} &\sum_{i=1}^{n}(|t_0|+t_i+\varepsilon)\mu(V_i) - |t_0|\mu(\operatorname{spt}(f)) \\ &\leq \sum_{i=1}^{n}(|t_0|+t_i+\varepsilon)(\mu(E_i)+\varepsilon/n) - |t_0|\mu(\operatorname{spt}(f)) \\ &\leq |t_0|\overbrace{\sum_{i=1}^{n}\mu(E_i)}^{\mu(\operatorname{spt}(f))} + |t_0|\varepsilon + \sum_{i=1}^{n} t_i\mu(E_i) + \varepsilon(|t_0|+|b|) \\ &\sum_{i=1}^{n} t_i\frac{\varepsilon}{n} + \varepsilon\sum_{i=1}^{n}\mu(E_i) + \varepsilon^2 - |t_0|\mu(\operatorname{spt}(f)). \end{aligned}$$

The first and last terms cancel. Therefore this is no larger than

$$\begin{aligned} &(2|t_0|+|b|+\mu(\operatorname{spt}(f))+\varepsilon)\varepsilon \\ &+\sum_{i=1}^{n} t_{i-1}\mu(E_i) + \varepsilon\mu(\operatorname{spt}(f)) + \sum_{i=1}^{n}(|t_0|+|b|)\frac{\varepsilon}{n} \\ &\leq \int f d\mu \end{aligned}$$

$$\leq \int f d\mu + (2|t_0|+|b|+2\mu(\operatorname{spt}(f))+\varepsilon)\varepsilon + (|t_0|+|b|)\varepsilon$$

Since $\varepsilon > 0$ is arbitrary, $Lf \leq \int f d\mu$ for all $f \in C_c(\Omega)$, f real valued. Hence equality holds because $L(-f) \leq -\int f d\mu$ so $L(f) \geq \int f d\mu$. Thus $Lf = \int f d\mu$ for all $f \in C_c(\Omega)$. Just apply the result for real functions to the real and imaginary parts of f. This gives the existence part of the Riesz representation theorem.

It only remains to prove uniqueness. Suppose both μ_1 and μ_2 are measures on $\mathcal{S}$ satisfying the conclusions of the theorem. Then if K is compact and $V \supseteq K$, let $K \prec f \prec V$. Then

$$\mu_1(K) \leq \int f d\mu_1 = Lf = \int f d\mu_2 \leq \mu_2(V).$$

Thus, taking the inf for all $V \supseteq K$, $\mu_1(K) \leq \mu_2(K)$ for all K. Similarly, the inequality can be reversed and so it follows the two measures are equal on compact sets. By the assumption of inner regularity on open sets, the two measures are also equal on all open sets. By outer regularity, they are equal on all sets of $\mathcal{S}$. ■

Example 19.7.9 *Let $L(f)=\int_{-\infty}^{\infty} f(t)\,dt$ for all $f\in C_c(\mathbb{R})$ where this is just the ordinary Riemann integral. Then the resulting measure is known as one dimensional Lebesgue measure.*

Example 19.7.10 *Let $L(f)=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(\boldsymbol{x})\,dx_1\cdots dx_n$ for $f\in C_c(\mathbb{R}^n)$. Then the resulting measure is m_n, n dimensional Lebesgue measure.*

Here is a nice observation.

Proposition 19.7.11 *In Example 19.7.10 the order of integration is not important. The same functional is obtained in any order.*

Proof: Let $\mathrm{spt}(f)\subseteq[-R,R]^n$. It clearly suffices to show this for $n=2$. Then by the definition of the Riemann integral, $\int_{-R}^{R}\int_{-R}^{R} f(x,y)\,dxdy=$

$$\begin{aligned}\int_{-R}^{R}\sum_i\int_{x_i}^{x_{i+1}} f(s,y)\,dsdy &= \sum_{j=0}^{n-1}\sum_{i=0}^{n-1}\int_{y_i}^{y_{i+1}}\int_{x_i}^{x_{i+1}} f(s,t)\,dsdt\\ &= \sum_{j=0}^{n-1}\sum_{i=0}^{n-1}\int_{y_i}^{y_{i+1}} f(s_i,t)\,(x_{i+1}-x_i)\,dt\end{aligned}$$

$$=\sum_{j=0}^{n-1}\sum_{i=0}^{n-1} f(s_i,t_j)\,(x_{i+1}-x_i)\,(y_{j+1}-y_j)=\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} f(s_i,t_j)\,(x_{i+1}-x_i)\,(y_{j+1}-y_j)$$

where $-R=x_0<x_1<\cdots<x_n=R$ is a uniform partition of $[-R,R]$ with the y_i also giving a uniform partition of $[-R,R]$. Similar reasoning implies

$$\int_{-R}^{R}\int_{-R}^{R} f(x,y)\,dydx=\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} f(\hat{s}_i,\hat{t}_j)\,(x_{i+1}-x_i)\,(y_{j+1}-y_j)$$

Now $(s_i,t_j),(\hat{s}_i,\hat{t}_j)$ are both in $[x_i,x_{i+1}]\times[y_j,y_{j+1}]$. Thus, by uniform continuity, if n is large enough,

$$\left|f(s_i,t_j)-f(\hat{s}_i,\hat{t}_j)\right|<\frac{\varepsilon}{4R^2}$$

Then it follows that

$$\left|\int_{-R}^{R}\int_{-R}^{R} f(x,y)\,dxdy-\int_{-R}^{R}\int_{-R}^{R} f(x,y)\,dydx\right|$$

$$\begin{aligned}&\leq \sum_{i=0}^{n-1}\sum_{j=0}^{n-1}\left|f(\hat{s}_i,\hat{t}_j)-f(s_i,t_j)\right|(x_{i+1}-x_i)\,(y_{j+1}-y_j)\\ &\leq \sum_{i=0}^{n-1}\sum_{j=0}^{n-1}\frac{\varepsilon}{4R^2}\,(x_{i+1}-x_i)\,(y_{j+1}-y_j)=\varepsilon\end{aligned}$$

Since ε is arbitrary, this shows that the two iterated integrals are the same. In case $n>2$, you can do exactly the same argument using the mean value theorem for integrals and obtain the same result by a similar argument, or you could use this result on pairs of integrals. ■

19.8 Slicing Measures

I saw this material first in the book [17]. It can be presented as an application of the theory of differentiation of Radon measures and the Riesz representation theorem for positive linear functionals. It is an amazing theorem and can be used to understand conditional probability.

Let μ be a finite Radon measure. I will show here that a formula of the following form holds.

$$\mu(F) = \int_F d\mu = \int_{\mathbb{R}^n} \int_{\mathbb{R}^m} \mathcal{X}_F(\boldsymbol{x},\boldsymbol{y})\, d\nu_{\boldsymbol{x}}(y)\, d\alpha(x)$$

where $\alpha(E) = \mu(E \times \mathbb{R}^m)$. When this is done, the measures, $\nu_{\boldsymbol{x}}$, are probability measures called slicing measures, and this shows that an integral with respect to μ can be written as an iterated integral in terms of the measure α and the slicing measures $\nu_{\boldsymbol{x}}$. It looks to me a lot like going backwards in the construction of product measure. One starts with a measure μ, defined on the Cartesian product and produces α and a possibly infinite family of slicing measures from it whereas in the construction of product measure, one starts with two measures and obtains a new measure on a σ algebra of subsets of the Cartesian product of two spaces. However, in contrast to product measures, these slicing measures are dependent on $\boldsymbol{x}$. This can be tied to the concept of independence or not, of random variables. They end up independent when the slicing measures don't depend on other variables. First here are two technical lemmas.

Lemma 19.8.1 *The space $C_c(\mathbb{R}^m)$ with the norm*

$$\|f\| \equiv \sup\{|f(\boldsymbol{y})| : \boldsymbol{y} \in \mathbb{R}^m\}$$

is separable.

Proof: Let $\mathscr{D}_l$ consist of all functions which are of the form

$$\sum_{|\alpha| \le N} a_\alpha \boldsymbol{y}^\alpha \left(\operatorname{dist}\left(\boldsymbol{y}, B(\boldsymbol{0}, l+1)^C\right)\right)^{n_\alpha}$$

where $a_\alpha \in \mathbb{Q}$, α is a multi-index, and n_α is a positive integer. Consider $\mathscr{D} \equiv \cup_l \mathscr{D}_l$. Then $\mathscr{D}$ is countable. If $f \in C_c(\mathbb{R}^n)$, then choose l large enough that $\operatorname{spt}(f) \subseteq B(\boldsymbol{0}, l+1)$, a locally compact space, $f \in C_0(B(\boldsymbol{0}, l+1))$. Then since $\mathscr{D}_l$ separates the points of $B(\boldsymbol{0}, l+1)$ is closed with respect to conjugates, and annihilates no point, it is dense in $C_0(B(\boldsymbol{0}, l+1))$ by the Stone Weierstrass theorem. Alternatively, $\mathscr{D}$ is dense in $C_0(\mathbb{R}^n)$ by Stone Weierstrass and $C_c(\mathbb{R}^n)$ is a subspace so it is also separable. So is $C_c(\mathbb{R}^n)^+$, the nonnegative functions in $C_c(\mathbb{R}^n)$. ■

From the regularity of Radon measures, the following lemma follows.

Lemma 19.8.2 *If μ and ν are two Radon measures defined on σ algebras, $\mathscr{S}_\mu$ and $\mathscr{S}_\nu$, of subsets of $\mathbb{R}^n$ and if $\mu(V) = \nu(V)$ for all V open, then $\mu = \nu$ and $\mathscr{S}_\mu = \mathscr{S}_\nu$.*

Proof: Since the two measures agree on open sets, it follows that the two measures agree on every G_δ and F_σ set. By Proposition 11.1.2 on Page 319, if $E \in \mathscr{S}_\mu$, then there exists F, G such that $F \subseteq E \subseteq G$ and $\mu(G) = \nu(G) = \nu(F) = \mu(F)$ with F an F_σ set and G a G_δ set. Then, by completeness of ν, $E \in \mathscr{S}_\nu$. Similarly $\mathscr{S}_\nu \subseteq \mathscr{S}_\mu$ so the two σ algebras are equal. Now letting F and G be as above,

$$\mu(E) \le \mu(G) = \nu(G) = \nu(E) = \nu(F) = \mu(F) \le \mu(E)$$

and so all inequalities are equalities and $\nu(E) = \mu(E)$ for all $E \in \mathcal{S} \equiv \mathcal{S}_\mu \equiv \mathcal{S}_\nu$. ■

The main result in the section is the following theorem.

Theorem 19.8.3 *Let μ be a finite Radon measure on $\mathbb{R}^{n+m}$ defined on a σ algebra, $\mathcal{F}$. Then there exists a unique finite Radon measure α, defined on a σ algebra $\mathcal{S}$, of sets of $\mathbb{R}^n$ which satisfies*

$$\alpha(E) = \mu(E \times \mathbb{R}^m) \tag{19.12}$$

for all E Borel. There also exists a Borel set of α measure zero N, such that for each $\boldsymbol{x} \notin N$, there exists a Radon probability measure $\nu_{\boldsymbol{x}}$ such that if f is a nonnegative μ measurable function or a μ measurable function in $L^1(\mu)$,

$$\boldsymbol{y} \to f(\boldsymbol{x},\boldsymbol{y}) \text{ is } \nu_{\boldsymbol{x}} \text{ measurable } \alpha \text{ a.e.}$$

$$\boldsymbol{x} \to \int_{\mathbb{R}^m} f(\boldsymbol{x},\boldsymbol{y})\, d\nu_{\boldsymbol{x}}(y) \text{ is } \alpha \text{ measurable} \tag{19.13}$$

and

$$\int_{\mathbb{R}^{n+m}} f(\boldsymbol{x},\boldsymbol{y})\, d\mu = \int_{\mathbb{R}^n} \left(\int_{\mathbb{R}^m} f(\boldsymbol{x},\boldsymbol{y})\, d\nu_{\boldsymbol{x}}(y) \right) d\alpha(x). \tag{19.14}$$

If $\widehat{\nu}_{\boldsymbol{x}}$ is any other collection of Radon measures satisfying 19.13 and 19.14, then $\widehat{\nu}_{\boldsymbol{x}} = \nu_{\boldsymbol{x}}$ for α a.e. $\boldsymbol{x}$.

Proof:

Existence and uniqueness of α

First consider the uniqueness of α. Suppose α_1 is another Radon measure satisfying 19.12. Then in particular, α_1 and α agree on open sets and so the two measures are the same by Lemma 19.8.2.

To establish the existence of α, define α_0 on Borel sets by $\alpha_0(E) = \mu(E \times \mathbb{R}^m)$ Thus α_0 is a finite Borel measure and so it is finite on compact sets. Lemma 20.2.3 on Page 548 implies the existence of the Radon measure α extending α_0.

Uniqueness of $\nu_{\boldsymbol{x}}$

Next consider the uniqueness of $\nu_{\boldsymbol{x}}$. Suppose $\nu_{\boldsymbol{x}}$ and $\widehat{\nu}_{\boldsymbol{x}}$ satisfy all conclusions of the theorem with exceptional sets denoted by N and $\widehat{N}$ respectively. Then, enlarging N and $\widehat{N}$, one may also assume, using Lemma 11.3.1, that for $\boldsymbol{x} \notin N \cup \widehat{N}$, $\alpha(B(\boldsymbol{x},r)) > 0$ whenever $r > 0$. Now let $A = \prod_{i=1}^m (a_i, b_i]$ where a_i and b_i are rational. Thus there are countably many such sets. Then from the conclusion of the theorem, if $\boldsymbol{x}_0 \notin N \cup \widehat{N}$,

$$\begin{aligned} & \frac{1}{\alpha(B(\boldsymbol{x}_0,r))} \int_{B(\boldsymbol{x}_0,r)} \int_{\mathbb{R}^m} \mathcal{X}_A(\boldsymbol{y})\, d\nu_{\boldsymbol{x}}(y)\, d\alpha \\ = & \frac{1}{\alpha(B(\boldsymbol{x}_0,r))} \int_{B(\boldsymbol{x}_0,r)} \int_{\mathbb{R}^m} \mathcal{X}_A(\boldsymbol{y})\, d\widehat{\nu}_{\boldsymbol{x}}(y)\, d\alpha, \end{aligned}$$

and by the Lebesgue Besicovitch Differentiation theorem, there exists a set of α measure zero, E_A, such that if $\boldsymbol{x}_0 \notin E_A \cup N \cup \widehat{N}$, then the limit in the above exists as $r \to 0$ and yields $\nu_{\boldsymbol{x}_0}(A) = \widehat{\nu}_{\boldsymbol{x}_0}(A)$. Letting E denote the union of all the sets E_A for A as described above, it follows that E is a set of measure zero and if $\boldsymbol{x}_0 \notin E \cup N \cup \widehat{N}$ then $\nu_{\boldsymbol{x}_0}(A) = \widehat{\nu}_{\boldsymbol{x}_0}(A)$ for

all such sets A. But every open set can be written as a disjoint union of sets of this form and so for all such $\boldsymbol{x}_0$, $\nu_{\boldsymbol{x}_0}(V) = \widehat{\nu}_{\boldsymbol{x}_0}(V)$ for all V open. By Lemma 19.8.2 this shows the two measures are equal and proves the uniqueness assertion for $\nu_{\boldsymbol{x}}$. It remains to show the existence of the measures $\nu_{\boldsymbol{x}}$.

Existence of $\nu_{\boldsymbol{x}}$

For $f \geq 0$, $f \in C_c(\mathbb{R}^m)$ and $g \in C_c(\mathbb{R}^n)$, define $g \to \int_{\mathbb{R}^{n+m}} g(\boldsymbol{x}) f(\boldsymbol{y}) d\mu$. Since $f \geq 0$, this is a positive linear functional on $C_c(\mathbb{R}^n)$. Therefore, there exists a unique Radon measure ν_f such that for all $g \in C_c(\mathbb{R}^n)$,

$$\int_{\mathbb{R}^{n+m}} g(\boldsymbol{x}) f(\boldsymbol{y}) d\mu = \int_{\mathbb{R}^n} g(\boldsymbol{x}) d\nu_f.$$

I claim that $\nu_f \ll \alpha$, the two being considered as measures on $\mathscr{B}(\mathbb{R}^n)$. Suppose then that K is a compact set and $\alpha(K) = 0$. Then let $K \prec g \prec V$ where V is open.

$$\nu_f(K) = \int_{\mathbb{R}^n} \mathscr{X}_K(\boldsymbol{x}) d\nu_f(\boldsymbol{x}) \leq \int_{\mathbb{R}^n} g(\boldsymbol{x}) d\nu_f(\boldsymbol{x}) = \int_{\mathbb{R}^{n+m}} g(\boldsymbol{x}) f(\boldsymbol{y}) d\mu$$

$$\leq \int_{\mathbb{R}^{m+m}} \mathscr{X}_{V\times\mathbb{R}^m}(\boldsymbol{x},\boldsymbol{y}) f(\boldsymbol{y}) d\mu \leq \|f\|_\infty \mu(V \times \mathbb{R}^m) = \|f\|_\infty \alpha(V)$$

Then for any $\varepsilon > 0$, one can choose V such that the right side is less than ε. Therefore, $\nu_f(K) = 0$ also. By regularity considerations, $\nu_f \ll \alpha$ as claimed.

It follows from the Radon Nikodym theorem, either Theorem 11.6.2 or the more abstract Radon Nikodym theorem which is presented earlier Theorem 10.13.7 on Page 309, there exists a function $h_f \in L^1(\alpha)$ such that for all $g \in C_c(\mathbb{R}^n)$,

$$\int_{\mathbb{R}^{n+m}} g(\boldsymbol{x}) f(\boldsymbol{y}) d\mu = \int_{\mathbb{R}^n} g(\boldsymbol{x}) d\nu_f = \int_{\mathbb{R}^n} g(\boldsymbol{x}) h_f(\boldsymbol{x}) d\alpha. \tag{19.15}$$

It is obvious from the formula that the map from $f \in C_c(\mathbb{R}^m)$ to $L^1(\alpha)$ given by $f \to h_f$ is linear in the sense that for $a, b \geq 0$ and $f, \hat{f} \geq 0, h_{af+b\hat{f}} = ah_f + bh_{\hat{f}}$. However, this is not sufficiently specific because functions in $L^1(\alpha)$ are only determined $a.e.$ However, for $h_f \in L^1(\alpha)$, you can specify a particular representative α a.e. By the fundamental theorem of calculus,

$$\widehat{h_f}(\boldsymbol{x}) \equiv \lim_{r\to 0} \frac{1}{\alpha(B(\boldsymbol{x},r))} \int_{B(\boldsymbol{x},r)} h_f(\boldsymbol{z}) d\alpha(z) \tag{19.16}$$

which exists off some set of α measure zero Z_f. Note that since this involves the integral over a ball, it does not matter which representative of h_f is placed in the formula. Therefore, $\widehat{h_f}(\boldsymbol{x})$ is well defined pointwise for all $\boldsymbol{x}$ not in some set of measure zero Z_f. Since $\widehat{h_f} = h_f$ a.e. it follows that $\widehat{h_f}$ is well defined and will work in the formula 19.15. Let $Z = \cup\{Z_f : f \in \mathscr{D}\}$ where $\mathscr{D}$ is a countable dense subset of $C_c(\mathbb{R}^m)^+$. Of course it is desired to have the limit 19.16 hold for all f, not just $f \in \mathscr{D}$. We will show that this limit 19.16 holds for all $\boldsymbol{x} \notin Z$. Thus, we will have $\boldsymbol{x} \to \widehat{h_f}(\boldsymbol{x})$ defined by the above limit off Z and so, since $\widehat{h_f}(\boldsymbol{x}) = h_f(\boldsymbol{x})$ a.e., it follows that for all $f \in C_c(\mathbb{R}^m)^+$

$$\int_{\mathbb{R}^{n+m}} g(\boldsymbol{x}) f(\boldsymbol{y}) d\mu = \int_{\mathbb{R}^n} g(\boldsymbol{x}) \widehat{h_f}(\boldsymbol{x}) d\alpha$$

One could then take $\widehat{h_f}(\boldsymbol{x})$ to be defined as 0 for $\boldsymbol{x} \notin Z$.

For f an arbitrary function in $C_c(\mathbb{R}^m)^+$ and $f' \in \mathscr{D}$, a dense countable subset of $C_c(\mathbb{R}^m)^+$, it follows from 19.15,

$$\begin{aligned}\left|\int_{\mathbb{R}^n} g(\boldsymbol{x})\left(h_f(\boldsymbol{x})-h_{f'}(\boldsymbol{x})\right)d\alpha\right| &= \left|\int_{\mathbb{R}^{n+m}} g(\boldsymbol{x})\left(f(\boldsymbol{y})-f'(\boldsymbol{y})\right)d\mu\right| \\ &\leq \left\|f-f'\right\|_\infty \int_{\mathbb{R}^{n+m}} |g(\boldsymbol{x})|\,d\mu\end{aligned}$$

Let $g_k(\boldsymbol{x}) \uparrow \mathscr{X}_{B(\boldsymbol{z},r)}(\boldsymbol{x})$ where $\boldsymbol{z} \notin Z$. Then by the dominated convergence theorem, the above implies

$$\left|\int_{B(\boldsymbol{z},r)}\left(h_f(\boldsymbol{x})-h_{f'}(\boldsymbol{x})\right)d\alpha\right| \leq \left\|f-f'\right\|_\infty \int_{B(\boldsymbol{z},r)\times\mathbb{R}^m} d\mu = \left\|f-f'\right\|_\infty \alpha(B(\boldsymbol{z},r)).$$

Dividing by $\alpha(B(\boldsymbol{z},r))$, it follows that if $\alpha(B(\boldsymbol{z},r)) > 0$ for all $r > 0$, then for all $r > 0$,

$$\left|\frac{1}{\alpha(B(\boldsymbol{z},r))}\int_{B(\boldsymbol{z},r)}\left(h_f(\boldsymbol{x})-h_{f'}(\boldsymbol{x})\right)d\alpha\right| \leq \left\|f-f'\right\|_\infty$$

It follows that for $f \in C_c(\mathbb{R}^m)^+$ arbitrary and $\boldsymbol{z} \notin Z$,

$$\begin{aligned}&\limsup_{r\to 0}\frac{1}{\alpha(B(\boldsymbol{z},r))}\int_{B(\boldsymbol{z},r)} h_f(\boldsymbol{x})\,d\alpha - \liminf_{r\to 0}\frac{1}{\alpha(B(\boldsymbol{z},r))}\int_{B(\boldsymbol{z},r)} h_f(\boldsymbol{x})\,d\alpha \\ =\ &\limsup_{r\to 0}\frac{1}{\alpha(B(\boldsymbol{z},r))}\int_{B(\boldsymbol{z},r)}\left(h_f(\boldsymbol{x})-h_{f'}(\boldsymbol{x})\right)d\alpha(x) \\ &-\liminf_{r\to 0}\frac{1}{\alpha(B(\boldsymbol{z},r))}\int_{B(\boldsymbol{z},r)}\left(h_f(\boldsymbol{x})-h_{f'}(\boldsymbol{x})\right)d\alpha(x) \\ \leq\ &\left|\limsup_{r\to 0}\frac{1}{\alpha(B(\boldsymbol{z},r))}\int_{B(\boldsymbol{z},r)}\left(h_f(\boldsymbol{x})-h_{f'}(\boldsymbol{x})\right)d\alpha(x)\right| \\ &+\left|\liminf_{r\to 0}\frac{1}{\alpha(B(\boldsymbol{z},r))}\int_{B(\boldsymbol{z},r)}\left(h_f(\boldsymbol{x})-h_{f'}(\boldsymbol{x})\right)d\alpha(x)\right| \leq 2\left\|f-f'\right\|_\infty\end{aligned}$$

and since f' is arbitrary, it follows that the limit of 19.16 holds for all $f \in C_c(\mathbb{R}^m)^+$ whenever $\boldsymbol{z} \notin Z$, the above set of measure zero.

Now for f an arbitrary real valued function of $C_c(\mathbb{R}^n)$, simply apply the above result to positive and negative parts to obtain $h_f \equiv h_{f^+} - h_{f^-}$ and $\widehat{h_f} \equiv \widehat{h_{f^+}} - \widehat{h_{f^-}}$. Then it follows that for all $f \in C_c(\mathbb{R}^m)$ and $g \in C_c(\mathbb{R}^m)$

$$\int_{\mathbb{R}^{n+m}} g(\boldsymbol{x})\,f(\boldsymbol{y})\,d\mu = \int_{\mathbb{R}^n} g(\boldsymbol{x})\,\widehat{h_f}(\boldsymbol{x})\,d\alpha.$$

It is obvious from 19.16 and 19.15 that for each $\boldsymbol{x} \notin Z$, the set of measure zero given above, that $f \to \widehat{h_f}(\boldsymbol{x})$ is a positive linear functional. (It is clear that it acts like a linear map for nonnegative f and so consideration of positive and negative parts of real and imaginary parts show this delivers a positive linear functional.) Hence by the Riesz representation theorem, there exists a unique $\nu_{\boldsymbol{x}}$ such that for all $\boldsymbol{x}$

$$\widehat{h_f}(\boldsymbol{x}) = \int_{\mathbb{R}^m} f(\boldsymbol{y})\,d\nu_{\boldsymbol{x}}(y).$$

It follows that

$$\int_{\mathbb{R}^{n+m}} g(\boldsymbol{x}) f(\boldsymbol{y}) d\mu = \int_{\mathbb{R}^n} \int_{\mathbb{R}^m} g(\boldsymbol{x}) f(\boldsymbol{y}) d\nu_{\boldsymbol{x}}(y) d\alpha(x) \tag{19.17}$$

and $\boldsymbol{x} \to \int_{\mathbb{R}^m} f(\boldsymbol{y}) d\nu_{\boldsymbol{x}}$ is α measurable and $\nu_{\boldsymbol{x}}$ is a Radon measure.

Now let $f_k \uparrow \mathcal{X}_{\mathbb{R}^m}$ and $g \geq 0$. Then by monotone convergence theorem,

$$\int_{\mathbb{R}^{n+m}} g(\boldsymbol{x}) d\mu = \int_{\mathbb{R}^n} g(\boldsymbol{x}) \int_{\mathbb{R}^m} d\nu_{\boldsymbol{x}} d\alpha$$

If $g_k \uparrow \mathcal{X}_{\mathbb{R}^n}$, the monotone convergence theorem shows that $\boldsymbol{x} \to \int_{\mathbb{R}^m} d\nu_{\boldsymbol{x}}$ is $L^1(\alpha)$.

Next let $g_k \uparrow \mathcal{X}_{B(\boldsymbol{x},r)}$ and use monotone convergence theorem to write

$$\alpha(B(\boldsymbol{x},r)) \equiv \int_{B(\boldsymbol{x},r)\times\mathbb{R}^m} d\mu = \int_{B(\boldsymbol{x},r)} \int_{\mathbb{R}^m} d\nu_{\boldsymbol{x}} d\alpha$$

Then dividing by $\alpha(B(\boldsymbol{x},r))$ and taking a limit as $r \to 0$, it follows that for α a.e. $\boldsymbol{x}$, $1 = \nu_{\boldsymbol{x}}(\mathbb{R}^m)$, so these $\nu_{\boldsymbol{x}}$ are probability measures off a set of α measure zero. Letting $g_k(\boldsymbol{x}) \uparrow \mathcal{X}_A(\boldsymbol{x})$, $f_k(\boldsymbol{y}) \uparrow \mathcal{X}_B(\boldsymbol{y})$ for A, B open, it follows that 19.17 is valid for $g(\boldsymbol{x})$ replaced with $\mathcal{X}_A(\boldsymbol{x})$ and $f(\boldsymbol{y})$ replaced with $\mathcal{X}_B(\boldsymbol{y})$.

Now let $\mathcal{G}$ denote the Borel sets F of $\mathbb{R}^{n+m}$ such that

$$\int_{\mathbb{R}^{n+m}} \mathcal{X}_F(\boldsymbol{x},\boldsymbol{y}) d\mu(x,y) = \int_{\mathbb{R}^n} \int_{\mathbb{R}^m} \mathcal{X}_F(\boldsymbol{x},\boldsymbol{y}) d\nu_{\boldsymbol{x}}(y) d\alpha(x)$$

and that all the integrals make sense. As just explained, this includes all Borel sets of the form $F = A \times B$ where A, B are open. It is clear that $\mathcal{G}$ is closed with respect to countable disjoint unions and complements, while sets of the form $A \times B$ for A, B open form a π system. Therefore, by Lemma 9.3.2 on Page 247, $\mathcal{G}$ contains the Borel sets which is the smallest σ algebra which contains such products of open sets. It follows from the usual approximation with simple functions that if $f \geq 0$ and is Borel measurable, then

$$\int_{\mathbb{R}^{n+m}} f(\boldsymbol{x},\boldsymbol{y}) d\mu(x,y) = \int_{\mathbb{R}^n} \int_{\mathbb{R}^m} f(\boldsymbol{x},\boldsymbol{y}) d\nu_{\boldsymbol{x}}(y) d\alpha(x)$$

with all the integrals making sense.

This proves the theorem in the case where f is Borel measurable and nonnegative. It just remains to extend this to the case where f is only μ measurable. However, from regularity of μ there exist Borel measurable functions $g, h, g \leq f \leq h$ such that

$$\int_{\mathbb{R}^{n+m}} f(\boldsymbol{x},\boldsymbol{y}) d\mu(x,y) = \int_{\mathbb{R}^{n+m}} g(\boldsymbol{x},\boldsymbol{y}) d\mu(x,y) = \int_{\mathbb{R}^{n+m}} h(\boldsymbol{x},\boldsymbol{y}) d\mu(x,y)$$

It follows

$$\int_{\mathbb{R}^n} \int_{\mathbb{R}^m} g(\boldsymbol{x},\boldsymbol{y}) d\nu_{\boldsymbol{x}}(y) d\alpha(x) = \int_{\mathbb{R}^n} \int_{\mathbb{R}^m} h(\boldsymbol{x},\boldsymbol{y}) d\nu_{\boldsymbol{x}}(y) d\alpha(x)$$

and so, since for α a.e. $\boldsymbol{x}, \boldsymbol{y} \to \boldsymbol{g}(\boldsymbol{x},\boldsymbol{y})$ and $\boldsymbol{y} \to \boldsymbol{h}(\boldsymbol{x},\boldsymbol{y})$ are $\nu_{\boldsymbol{x}}$ measurable with

$$0 = \int_{\mathbb{R}^m} (h(\boldsymbol{x},\boldsymbol{y}) - g(\boldsymbol{x},\boldsymbol{y})) d\nu_{\boldsymbol{x}}(y)$$

and $\nu_{\boldsymbol{x}}$ is a Radon measure, hence complete, it follows for α a.e. $\boldsymbol{x}$, $\boldsymbol{y} \to f(\boldsymbol{x},\boldsymbol{y})$ must be $\nu_{\boldsymbol{x}}$ measurable because it is equal to $\boldsymbol{y} \to g(\boldsymbol{x},\boldsymbol{y})$, $\nu_{\boldsymbol{x}}$ a.e. Therefore, for α a.e. $\boldsymbol{x}$, it

makes sense to write $\int_{\mathbb{R}^m} f(\boldsymbol{x},\boldsymbol{y})\,d\nu_{\boldsymbol{x}}(y)$. Similar reasoning applies to the above function of $\boldsymbol{x}$ being α measurable due to α being complete. It follows

$$\int_{\mathbb{R}^{n+m}} f(\boldsymbol{x},\boldsymbol{y})\,d\mu(x,y) = \int_{\mathbb{R}^{n+m}} g(\boldsymbol{x},\boldsymbol{y})\,d\mu(x,y)$$

$$= \int_{\mathbb{R}^n}\int_{\mathbb{R}^m} g(\boldsymbol{x},\boldsymbol{y})\,d\nu_{\boldsymbol{x}}(y)\,d\alpha(x) = \int_{\mathbb{R}^n}\int_{\mathbb{R}^m} f(\boldsymbol{x},\boldsymbol{y})\,d\nu_{\boldsymbol{x}}(y)\,d\alpha(x)$$

with everything making sense. ■

19.9 Exercises

1. Let X be a finite dimensional normed linear space, real or complex. Show that X is separable. **Hint:** Let $\{v_i\}_{i=1}^n$ be a basis and define a map from $\mathbb{F}^n$ to X, θ, as follows. $\theta(\sum_{k=1}^n x_k \boldsymbol{e}_k) \equiv \sum_{k=1}^n x_k v_k$. Show θ is continuous and has a continuous inverse. Now let D be a countable dense set in $\mathbb{F}^n$ and consider $\theta(D)$.

2. Let $\alpha \in (0,1]$. We define, for X a compact subset of $\mathbb{R}^p$,

$$C^\alpha(X;\mathbb{R}^n) \equiv \{\boldsymbol{f} \in C(X;\mathbb{R}^n) : \rho_\alpha(\boldsymbol{f}) + \|\boldsymbol{f}\| \equiv \|\boldsymbol{f}\|_\alpha < \infty\}$$

where $\|\boldsymbol{f}\| \equiv \sup\{|\boldsymbol{f}(\boldsymbol{x})| : \boldsymbol{x} \in X\}$ and

$$\rho_\alpha(\boldsymbol{f}) \equiv \sup\{\frac{|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{y})|}{|\boldsymbol{x}-\boldsymbol{y}|^\alpha} : \boldsymbol{x},\boldsymbol{y} \in X,\ \boldsymbol{x} \neq \boldsymbol{y}\}.$$

Show that $(C^\alpha(X;\mathbb{R}^n), \|\cdot\|_\alpha)$ is a complete normed linear space. This is called a Holder space. What would this space consist of if $\alpha > 1$?

3. Let $\{\boldsymbol{f}_n\}_{n=1}^\infty \subseteq C^\alpha(X;\mathbb{R}^n)$ where X is a compact subset of $\mathbb{R}^p$ and suppose

$$\|\boldsymbol{f}_n\|_\alpha \leq M$$

for all n. Show there exists a subsequence, n_k, such that $\boldsymbol{f}_{n_k}$ converges in $C(X;\mathbb{R}^n)$. We say the given sequence is precompact when this happens. (This also shows the embedding of $C^\alpha(X;\mathbb{R}^n)$ into $C(X;\mathbb{R}^n)$ is a compact embedding.) **Hint:** You might want to use the Ascoli Arzela theorem.

4. Let $\boldsymbol{f} : \mathbb{R}\times\mathbb{R}^n \to \mathbb{R}^n$ be continuous and bounded and let $\boldsymbol{x}_0 \in \mathbb{R}^n$. If $\boldsymbol{x} : [0,T] \to \mathbb{R}^n$ and $h > 0$, let

$$\tau_h \boldsymbol{x}(s) \equiv \begin{cases} \boldsymbol{x}_0 \text{ if } s \leq h, \\ \boldsymbol{x}(s-h), \text{ if } s > h. \end{cases}$$

For $t \in [0,T]$, let $\boldsymbol{x}_h(t) = \boldsymbol{x}_0 + \int_0^t \boldsymbol{f}(s,\tau_h\boldsymbol{x}_h(s))\,ds$.Show using the Ascoli Arzela theorem that there exists a sequence $h \to 0$ such that $\boldsymbol{x}_h \to \boldsymbol{x}$ in $C([0,T];\mathbb{R}^n)$. Next argue $\boldsymbol{x}(t) = \boldsymbol{x}_0 + \int_0^t \boldsymbol{f}(s,\boldsymbol{x}(s))\,ds$ and conclude the following theorem. If $\boldsymbol{f} : \mathbb{R}\times\mathbb{R}^n \to \mathbb{R}^n$ is continuous and bounded, and if $\boldsymbol{x}_0 \in \mathbb{R}^n$ is given, there exists a solution to the following initial value problem.

$$\boldsymbol{x}' = \boldsymbol{f}(t,\boldsymbol{x}),\ t \in [0,T],\ \boldsymbol{x}(0) = \boldsymbol{x}_0.$$

This is the Peano existence theorem for ordinary differential equations.

5. Suppose $f \in C_0([0,\infty))$ and also $|f(t)| \leq Ce^{-rt}$. Let $\mathscr{A}$ denote the algebra of linear combinations of functions of the form e^{-st} for s sufficiently large. Thus $\mathscr{A}$ is dense in $C_0([0,\infty))$. Show that if $\int_0^\infty e^{-st} f(t)\,dt = 0$ for each s sufficiently large, then $f(t) = 0$. Next consider only $|f(t)| \leq Ce^{rt}$ for some r. That is f has exponential growth. Show the same conclusion holds for f if $\int_0^\infty e^{-st} f(t)\,dt = 0$ for all s sufficiently large. This justifies the Laplace transform procedure of differential equations where if the Laplace transforms of two functions are equal, then the two functions are considered to be equal. More can be said about this. **Hint:** For the last part, consider $g(t) \equiv e^{-2rt} f(t)$ and apply the first part to g. If $g(t) = 0$ then so is $f(t)$.

6. A set S along with an order $\leq$ is said to be a well ordered set if every nonempty subset of S has a smallest element. Here $\leq$ is an order in the usual way. If $x, y \in S$, then either $x \leq y$ or $y \leq x$ and it satisfies the transitive law: If $x \leq y$ and $y \leq z$, then $x \leq z$. Using the Hausdorff maximal theorem, show that every nonempty set can be well ordered. That is, there is an order for which the given set is well ordered. In particular $\mathbb{Q}$ the rational numbers can be well ordered. However, show that there is no way that the well order can coincide with the usual order on any open interval.

7. Verify that $P(\mathbb{N})$ the set of all subsets of the natural numbers is uncountable. Thus there exist uncountable sets. Pick an uncountable set Ω. Then you can consider this set to be well ordered, with the order denoted as $\leq$. Consider $\hat{\Omega} \equiv$

   $$\{\omega \in \Omega \text{ such that there are uncountably many elements of } \Omega \text{ less than } \omega\}$$

   If $\hat{\Omega} = \emptyset$, let $\Omega_0 = \hat{\Omega}$. If $\hat{\Omega} \neq \emptyset$, let ω_0 be the first element of $\hat{\Omega}$ and in this case let $\Omega_0 \equiv \{\omega : \omega < \omega_0\}$ That is $\omega \neq \omega_0$ and $\omega \leq \omega_0$. Explain why Ω_0 is uncountable. Explain why every element of Ω_0 has a "next" element. Now define a topology in the usual way. In particular, show that sets of the form $[\alpha, b), (a, b)$ where α is the first element of Ω_0 is a basis for a topology for Ω_0. Verify that a sub-basis is sets of the form $[\alpha, b), (a, \omega_0)$. Explain why this is a Hausdorff space. Explain why every element of Ω_0 is preceeded by countably many elements of Ω_0 and show every increasing sequence converges. Show that this cannot be a separable space. Suppose you have a cover of $(a,b]$ consisting of "sub-basic" open sets. Without loss of generality, all of these have nonempty intersection with (a,b). Let p be the first such that (p, ω_0) is in the open cover, assuming there are such sets. Thus you could simply use (p, ω_0) instead of all the others. If $p \leq a$, you are done. If not, then $p \in (a,b]$ and so some set of the other kind, $[\alpha, q)$ must contain p and so at most two sets from the open cover contain $\overline{(a,b)}$. Consider the other cases to verify that $(a,b]$ is compact. Now explain why $\overline{(a,b)}$ is either equal to $(a,b]$ or (a,b). In the second case, verify that (a,b) would be of the form $(a,\hat{b}]$ which was just shown to be compact thanks to Alexander sub-basis theorem. Is Ω_0 locally compact?

8. In the above example, show that Ω_0 is not compact. However, show that every sequence has a subsequence which converges. Recall that in any metric space, compactness and sequential compactness are equivalent. Hence one can conclude that there is no metric which will deliver the same topology for Ω_0 described above. That is to say, this horrible topological space is not metrizable. However, the Riesz representation theorem presented above would hold for this terrible thing.

Chapter 20

Product Measures

Sometimes it is necessary to consider infinite Cartesian products of topological spaces.

20.1 Algebras

First of all, here is the definition of an algebra and theorems which tell how to recognize one when you see it. An algebra is like a σ algebra except it is only closed with respect to finite unions.

Definition 20.1.1 *$\mathscr{A}$ is said to be an algebra of subsets of a set Z if $Z \in \mathscr{A}$, $\emptyset \in \mathscr{A}$, and when $E, F \in \mathscr{A}$, $E \cup F$ and $E \setminus F$ are both in $\mathscr{A}$.*

It is important to note that if $\mathscr{A}$ is an algebra, then it is also closed under finite intersections. This is because $E \cap F = (E^C \cup F^C)^C \in \mathscr{A}$ since $E^C = Z \setminus E \in \mathscr{A}$ and $F^C = Z \setminus F \in \mathscr{A}$. Note that every σ algebra is an algebra but not the other way around.

Something satisfying the above definition is called an algebra because union is like addition, the set difference is like subtraction and intersection is like multiplication. Furthermore, only finitely many operations are done at a time and so there is nothing like a limit involved.

How can you recognize an algebra when you see one? The answer to this question is the purpose of the following lemma.

Lemma 20.1.2 *Suppose $\mathscr{R}$ and $\mathscr{E}$ are subsets of $\mathscr{P}(Z)$[1] such that $\mathscr{E}$ is defined as the set of all **finite disjoint** unions of sets of $\mathscr{R}$. Suppose also*

$$\emptyset, Z \in \mathscr{R}$$

$$A \cap B \in \mathscr{R} \text{ whenever } A, B \in \mathscr{R},$$

$$A \setminus B \in \mathscr{E} \text{ whenever } A, B \in \mathscr{R}.$$

Then $\mathscr{E}$ is an algebra of sets of Z.

Proof: Note first that if $A \in \mathscr{R}$, then $A^C \in \mathscr{E}$ because $A^C = Z \setminus A$.

Now suppose that E_1 and E_2 are in $\mathscr{E}$,

$$E_1 = \cup_{i=1}^m R_i, \ E_2 = \cup_{j=1}^n R_j$$

where the R_i are disjoint sets in $\mathscr{R}$ and the R_j are disjoint sets in $\mathscr{R}$. Then

$$E_1 \cap E_2 = \cup_{i=1}^m \cup_{j=1}^n R_i \cap R_j$$

which is clearly an element of $\mathscr{E}$ because no two of the sets in the union can intersect and by assumption they are all in $\mathscr{R}$. Thus by induction, finite intersections of sets of $\mathscr{E}$ are in $\mathscr{E}$. Consider the difference of two elements of $\mathscr{E}$ next.

If $E = \cup_{i=1}^n R_i \in \mathscr{E}$,

$$E^C = \cap_{i=1}^n R_i^C = \text{ finite intersection of sets of } \mathscr{E}$$

[1] Set of all subsets of Z

which was just shown to be in $\mathscr{E}$. Now, if $E_1, E_2 \in \mathscr{E}$,

$$E_1 \setminus E_2 = E_1 \cap E_2^C \in \mathscr{E}$$

from what was just shown about finite intersections.

Finally consider finite unions of sets of $\mathscr{E}$. Let E_1 and E_2 be sets of $\mathscr{E}$. Then

$$E_1 \cup E_2 = (E_1 \setminus E_2) \cup E_2 \in \mathscr{E}$$

because $E_1 \setminus E_2$ consists of a finite disjoint union of sets of $\mathscr{R}$ and these sets must be disjoint from the sets of $\mathscr{R}$ whose union yields E_2 because $(E_1 \setminus E_2) \cap E_2 = \emptyset$. ■

The following corollary is particularly helpful in verifying the conditions of the above lemma.

Corollary 20.1.3 *Let $(Z_1, \mathscr{R}_1, \mathscr{E}_1)$ and $(Z_2, \mathscr{R}_2, \mathscr{E}_2)$ be as described in Lemma 20.1.2. Then $(Z_1 \times Z_2, \mathscr{R}, \mathscr{E})$ also satisfies the conditions of Lemma 20.1.2 if $\mathscr{R}$ is defined as*

$$\mathscr{R} \equiv \{R_1 \times R_2 : R_i \in \mathscr{R}_i\}$$

and

$$\mathscr{E} \equiv \{ \textit{finite disjoint unions of sets of } \mathscr{R}\}.$$

Consequently, $\mathscr{E}$ is an algebra of sets.

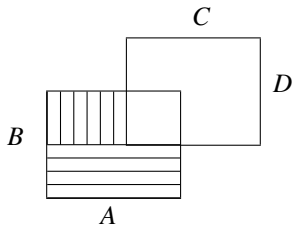
Proof: It is clear $\emptyset, Z_1 \times Z_2 \in \mathscr{R}$. Let $A \times B$ and $C \times D$ be two elements of $\mathscr{R}$.

$$A \times B \cap C \times D = (A \cap C) \times (B \cap D) \in \mathscr{R}$$

by assumption.

$$(A \times B) \setminus (C \times D) = A \times \overbrace{(B \setminus D)}^{\in \mathscr{E}_2} \cup \overbrace{(A \setminus C)}^{\in \mathscr{E}_1} \times \overbrace{(D \cap B)}^{\in \mathscr{R}_2} = (A \times Q) \cup (P \times R)$$

where $Q \in \mathscr{E}_2$, $P \in \mathscr{E}_1$, and $R \in \mathscr{R}_2$.



Since $A \times Q$ and $P \times R$ do not intersect, it follows that the above expression is in $\mathscr{E}$ because each of these terms are. ■

20.2 Caratheodory Extension Theorem

The Caratheodory extension theorem is a fundamental result which makes possible the consideration of measures on infinite products among other things. The idea is that if a finite measure defined only on an algebra is trying to be a measure, then in fact it can be extended to a measure.

Definition 20.2.1 *Let $\mathscr{E}$ be an algebra of sets of Ω. Thus $\mathscr{E}$ contains $\emptyset, \Omega$, and is closed with respect to differences and finite unions. Then μ_0 is a finite measure on $\mathscr{E}$ means μ_0 is finitely additive: **If** E_i, E are sets of $\mathscr{E}$ with the E_i disjoint and $E = \cup_{i=1}^{\infty} E_i$, then $\mu(E) = \sum_{i=1}^{\infty} \mu(E_i)$*

In this definition, μ_0 is trying to be a measure and acts like one whenever possible. Note the extra assumption that $E \in \mathscr{E}$. This would be automatic if it were a finite sum and finite union. Under these conditions, μ_0 can be extended uniquely to a complete measure μ, defined on a σ algebra of sets containing $\mathscr{E}$ such that μ agrees with μ_0 on $\mathscr{E}$. The following is the main result.

Theorem 20.2.2 *Let μ_0 be a measure on an algebra of sets $\mathscr{E}$, which satisfies $\mu_0(\Omega) < \infty$. Then there exists a complete measure space $(\Omega, \mathscr{S}, \mu)$ such that $\mu(E) = \mu_0(E)$ for all $E \in \mathscr{E}$. Also if ν is any measure which agrees with μ_0 on $\mathscr{E}$, then $\nu = \mu$ on $\sigma(\mathscr{E})$, the σ algebra generated by $\mathscr{E}$.*

Proof: Define an outer measure as follows.

$$\mu(S) \equiv \inf\left\{\sum_{i=1}^{\infty} \mu_0(E_i) : S \subseteq \cup_{i=1}^{\infty} E_i, E_i \in \mathscr{E}\right\}$$

Claim 1: μ is an outer measure.
Proof of Claim 1: Let $S \subseteq \cup_{i=1}^{\infty} S_i$ and let $S_i \subseteq \cup_{j=1}^{\infty} E_{ij}$, where

$$\mu(S_i) + \frac{\varepsilon}{2^i} \geq \sum_{j=1}^{\infty} \mu(E_{ij}).$$

Then

$$\mu(S) \leq \sum_i \sum_j \mu(E_{ij}) = \sum_i \left(\mu(S_i) + \frac{\varepsilon}{2^i}\right) = \sum_i \mu(S_i) + \varepsilon.$$

Since ε is arbitrary, this shows μ is an outer measure as claimed.

By the Caratheodory procedure, there exists a unique σ algebra $\mathscr{S}$, consisting of the μ measurable sets such that $(\Omega, \mathscr{S}, \mu)$ is a complete measure space. It remains to show that μ extends μ_0.

Claim 2: If $\mathscr{S}$ is the σ algebra of μ measurable sets, $\mathscr{S} \supseteq \mathscr{E}$ and $\mu = \mu_0$ on $\mathscr{E}$.

Proof of Claim 2: First observe that if $A \in \mathscr{E}$, then $\mu(A) \leq \mu_0(A)$ by definition. It remains to turn the inequality around. Letting

$$\mu(A) + \varepsilon > \sum_{i=1}^{\infty} \mu_0(E_i),\ \cup_{i=1}^{\infty} E_i \supseteq A,\ E_i \in \mathscr{E},$$

it follows that

$$\mu(A) + \varepsilon > \sum_{i=1}^{\infty} \mu_0(E_i \cap A) \geq \mu_0(A)$$

since $A = \cup_{i=1}^{\infty} E_i \cap A$. Therefore, $\mu = \mu_0$ on $\mathscr{E}$.

Consider the assertion that $\mathscr{E} \subseteq \mathscr{S}$. Let $A \in \mathscr{E}$ and let $S \subseteq \Omega$ be any set. By definition, there exist sets $\{E_i\} \subseteq \mathscr{E}$ such that $\cup_{i=1}^{\infty} E_i \supseteq S$ but

$$\mu(S) + \varepsilon > \sum_{i=1}^{\infty} \mu(E_i) = \sum_{i=1}^{\infty} \mu_0(E_i).$$

Then by the assumption that μ_0 is a measure on $\mathscr{E}$ and that $\mu = \mu_0$ on sets of $\mathscr{E}$,

$$\mu(S) \leq \mu(S \cap A) + \mu(S \setminus A)$$

$$\leq \mu\left(\left(\cup_{i=1}^{\infty} E_i\right) \setminus A\right) + \mu\left(\cup_{i=1}^{\infty} (E_i \cap A)\right) \leq \sum_{i=1}^{\infty} \mu(E_i \setminus A) + \sum_{i=1}^{\infty} \mu(E_i \cap A)$$

$$= \sum_{i=1}^{\infty} \mu_0(E_i \setminus A) + \sum_{i=1}^{\infty} \mu_0(E_i \cap A) = \sum_{i=1}^{\infty} \mu(E_i) < \mu(S) + \varepsilon.$$

Since ε is arbitrary, this shows $A \in \mathscr{S}$.

This has proved the existence part of the theorem. To verify uniqueness, $\sigma(\mathscr{E}) \subseteq \mathscr{S}$. Let

$$\mathscr{G} \equiv \{E \in \sigma(\mathscr{E}) : \mu(E) = \nu(E)\}.$$

Then $\mathscr{G}$ is given to contain $\mathscr{E}$ and is obviously closed with respect to countable disjoint unions and complements because $\sigma(\mathscr{E}) \subseteq \mathscr{S}$. Therefore by Dynkin's theorem, Lemma 9.3.2, $\mathscr{G} = \sigma(\mathscr{E})$. ■

The following lemma is also very significant. Actually Lemmas 9.4.4 and 9.4.5 are of even more use, but one can use the following to get useful information in some cases. In particular, in the proof of the Kolmogorov extension theorem, one can consider the special case that $M_t = \mathbb{R}$ or $\mathbb{R}^{n_t}$ in that theorem. The following lemma is Corollary 9.4.9 on Page 9.4.9. However, I am giving a different proof here to emphasize the theorem on positive linear functionals and measures.

Lemma 20.2.3 *Let M be a metric space with the closed balls compact and suppose μ is a measure defined on the Borel sets of M which is finite on compact sets. Then there exists a unique Radon measure, $\overline{\mu}$ which equals μ on the Borel sets. In particular μ must be both inner and outer regular on all Borel sets.*

Proof: Define a positive linear functional, $\Lambda(f) = \int f d\mu$. Let $\overline{\mu}$ be the Radon measure which comes from the Riesz representation theorem for positive linear functionals. Thus for all $f \in C_c(M)$,

$$\int f d\mu = \int f d\overline{\mu}.$$

If V is an open set, let $\{f_n\}$ be a sequence of continuous functions in $C_c(M)$ which is increasing and converges to $\mathscr{X}_V$ pointwise. Then applying the monotone convergence theorem,

$$\int \mathscr{X}_V d\mu = \mu(V) = \int \mathscr{X}_V d\overline{\mu} = \overline{\mu}(V)$$

and so the two measures coincide on all open sets. Every compact set is a countable intersection of open sets and so the two measures coincide on all compact sets. Now let $B(a,n)$ be a ball of radius n and let E be a Borel set contained in this ball. Then by regularity of $\overline{\mu}$ there exist sets F, G such that G is a countable intersection of open sets and F is a countable union of compact sets such that $F \subseteq E \subseteq G$ and $\overline{\mu}(G \setminus F) = 0$. Now $\mu(G) = \overline{\mu}(G)$ and $\mu(F) = \overline{\mu}(F)$. Thus

$$\begin{aligned} \overline{\mu}(G \setminus F) + \overline{\mu}(F) &= \overline{\mu}(G) \\ &= \mu(G) = \mu(G \setminus F) + \mu(F) \end{aligned}$$

and so $\mu(G \setminus F) = \overline{\mu}(G \setminus F)$. It follows

$$\mu(E) = \mu(F) = \overline{\mu}(F) = \overline{\mu}(G) = \overline{\mu}(E).$$

If E is an arbitrary Borel set, then $\mu(E \cap B(a,n)) = \overline{\mu}(E \cap B(a,n))$ and letting $n \to \infty$, this yields $\mu(E) = \overline{\mu}(E)$. ■

20.3 Kolmogorov Extension Theorem

This extension theorem is one of the most important theorems in probability theory, at least according to my understanding of the situation. As an example, one sometimes wants to consider infinitely many independent normally distributed random variables. Is there a probability space such that this kind of thing even exists? The answer is yes and one way to show this is through the use of the Kolmogorov extension theorem. I am presenting the most general version of this theorem that I have seen. For another proof see the book by Strook [55]. What I am using here is a modification of one in Billingsley [6].

Let M_t be a complete separable metric space. This is called a Polish space. I will denote a totally ordered index set, (Like $\mathbb{R}$) and the interest will be in building a measure on the product space, $\prod_{t\in I} M_t$. If you like less generality, just think of $M_t = \mathbb{R}^{k_t}$ or even $M_t = \mathbb{R}$. By the well ordering principle, you can always put an order on any index set so this order is no restriction, but we do not insist on a well order and in fact, index sets of most interest are $\mathbb{R}$ or $[0,\infty)$. Also for X a topological space, $\mathscr{B}(X)$ will denote the Borel sets.

Notation 20.3.1 *The symbol J will denote a finite subset of I, $J = (t_1, \cdots, t_n)$, the t_i taken in order. $\boldsymbol{E}_J$ will denote a set which has a set E_t of $\mathscr{B}(M_t)$ in the t^{th} position for $t \in J$ and for $t \notin J$, the set in the t^{th} position will be M_t. $\boldsymbol{K}_J$ will denote a set which has a compact set in the t^{th} position for $t \in J$ and for $t \notin J$, the set in the t^{th} position will be M_t. Also denote by $\mathscr{R}_J$ the sets $\boldsymbol{E}_J$ and $\mathscr{R}$ the union of all such $\mathscr{R}_J$. Let $\mathscr{E}_J$ denote finite disjoint unions of sets of $\mathscr{R}_J$ and let $\mathscr{E}$ denote finite disjoint unions of sets of $\mathscr{R}$. Thus if $\boldsymbol{F}$ is a set of $\mathscr{E}$, there exists J such that $\boldsymbol{F}$ is a finite disjoint union of sets of $\mathscr{R}_J$. For $\boldsymbol{F} \in \Omega$, denote by $\pi_J(\boldsymbol{F})$ the set $\prod_{t\in J} F_t$ where $\boldsymbol{F} = \prod_{t\in I} F_t$.*

Lemma 20.3.2 *The sets $\mathscr{E}$,$\mathscr{E}_J$ defined above form an algebra of sets of $\prod_{t\in I} M_t$.*

Proof: First consider $\mathscr{R}_J$. If $\boldsymbol{A}, \boldsymbol{B} \in \mathscr{R}_J$, then $\boldsymbol{A} \cap \boldsymbol{B} \in \mathscr{R}_J$ also. Is $\boldsymbol{A} \setminus \boldsymbol{B}$ a finite disjoint union of sets of $\mathscr{R}_J$? It suffices to verify that $\pi_J(\boldsymbol{A} \setminus \boldsymbol{B})$ is a finite disjoint union of $\pi_J(\mathscr{R}_J)$. Let $|J|$ denote the number of indices in J. If $|J| = 1$, then it is obvious that $\pi_J(\boldsymbol{A} \setminus \boldsymbol{B})$ is a finite disjoint union of sets of $\pi_J(\mathscr{R}_J)$. In fact, letting $J = (t)$ and the t^{th} entry of $\boldsymbol{A}$ is A and the t^{th} entry of $\boldsymbol{B}$ is B, then the t^{th} entry of $\boldsymbol{A} \setminus \boldsymbol{B}$ is $A \setminus B$, a Borel set of M_t, a finite disjoint union of Borel sets of M_t.

Suppose then that for $\boldsymbol{A}, \boldsymbol{B}$ sets of $\mathscr{R}_J$, $\pi_J(\boldsymbol{A} \setminus \boldsymbol{B})$ is a finite disjoint union of sets of $\pi_J(\mathscr{R}_J)$ for $|J| \leq n$, and consider $J = (t_1, \cdots, t_n, t_{n+1})$. Let the t_i^{th} entry of $\boldsymbol{A}$ and $\boldsymbol{B}$ be respectively A_i and B_i. It follows that $\pi_J(\boldsymbol{A} \setminus \boldsymbol{B})$ has the following in the entries for J

$$(A_1 \times A_2 \times \cdots \times A_n \times A_{n+1}) \setminus (B_1 \times B_2 \times \cdots \times B_n \times B_{n+1})$$

Letting A represent $A_1 \times A_2 \times \cdots \times A_n$ and B represent $B_1 \times B_2 \times \cdots \times B_n$, $\boldsymbol{A} \setminus \boldsymbol{B}$ is of the form

$$A \times (A_{n+1} \setminus B_{n+1}) \cup (A \setminus B) \times (A_{n+1} \cap B_{n+1})$$

By induction, $(A \setminus B)$ is the finite disjoint union of sets of $\mathscr{R}_{(t_1,\cdots,t_n)}$. Therefore, the above is the finite disjoint union of sets of $\mathscr{R}_J$. It follows that $\mathscr{E}_J$ is an algebra.

Now suppose $\boldsymbol{A}, \boldsymbol{B} \in \mathscr{R}$. Then for some finite set J, both are in $\mathscr{R}_J$. Then from what was just shown,

$$\boldsymbol{A} \setminus \boldsymbol{B} \in \mathscr{E}_J \subseteq \mathscr{E}, \boldsymbol{A} \cap \boldsymbol{B} \in \mathscr{R}.$$

By Lemma 20.1.2 on Page 545 this shows $\mathscr{E}$ is an algebra. ■

With this preparation, here is the Kolmogorov extension theorem. In the statement and proof of the theorem, F_i, G_i, and E_i will denote Borel sets. Any list of indices from I will always be assumed to be taken in order. Thus, if $J \subseteq I$ and $J = (t_1, \cdots, t_n)$, it will always be assumed $t_1 < t_2 < \cdots < t_n$.

Theorem 20.3.3 *For each finite set*

$$J = (t_1, \cdots, t_n) \subseteq I,$$

suppose there exists a Borel probability measure, $\nu_J = \nu_{t_1 \cdots t_n}$ defined on the Borel sets of $\prod_{t \in J} M_t$ such that the following consistency condition holds. If

$$(t_1, \cdots, t_n) \subseteq (s_1, \cdots, s_p),$$

then

$$\nu_{t_1 \cdots t_n}(F_{t_1} \times \cdots \times F_{t_n}) = \nu_{s_1 \cdots s_p}(G_{s_1} \times \cdots \times G_{s_p}) \tag{20.1}$$

where if $s_i = t_j$, then $G_{s_i} = F_{t_j}$ and if s_i is not equal to any of the indices t_k, then $G_{s_i} = M_{s_i}$. Then for $\mathscr{E}$ defined in Notation 20.3.1, there exists a probability measure P and a σ algebra $\mathscr{F} = \sigma(\mathscr{E})$ such that

$$\left(\prod_{t \in I} M_t, P, \mathscr{F} \right)$$

is a probability space. Also there exist measurable functions, $X_s : \prod_{t \in I} M_t \to M_s$ defined for $s \in I$ as $X_s \boldsymbol{x} \equiv x_s$ such that for each $(t_1 \cdots t_n) \subseteq I$,

$$\nu_{t_1 \cdots t_n}(F_{t_1} \times \cdots \times F_{t_n}) = P([X_{t_1} \in F_{t_1}] \cap \cdots \cap [X_{t_n} \in F_{t_n}])$$

$$= P\left((X_{t_1}, \cdots, X_{t_n}) \in \prod_{j=1}^{n} F_{t_j} \right) = P\left(\prod_{t \in I} F_t \right) \tag{20.2}$$

where $F_t = M_t$ for every $t \notin \{t_1 \cdots t_n\}$ and F_{t_i} is a Borel set. Also if f is a nonnegative function of finitely many variables, $x_{t_1}, \cdots, x_{t_n}$, measurable with respect to $\mathscr{B}\left(\prod_{j=1}^{n} M_{t_j}\right)$, then f is also measurable with respect to $\mathscr{F}$ and

$$\int_{M_{t_1} \times \cdots \times M_{t_n}} f(x_{t_1}, \cdots, x_{t_n}) \, d\nu_{t_1 \cdots t_n} = \int_{\prod_{t \in I} M_t} f(x_{t_1}, \cdots, x_{t_n}) \, dP \tag{20.3}$$

Proof: Let $\mathscr{E}$ be the algebra of sets defined in the above notation. I want to define a measure on $\mathscr{E}$. For $\boldsymbol{F} \in \mathscr{E}$, there exists J such that $\boldsymbol{F}$ is the finite disjoint union of sets of $\mathscr{R}_J$. Define

$$P_0(\boldsymbol{F}) \equiv \nu_J(\pi_J(\boldsymbol{F}))$$

Then P_0 is well defined because of the consistency condition on the measures ν_J. P_0 is clearly finitely additive because the ν_J are measures and one can pick J as large as desired to include all t where there may be something other than M_t. Also, from the definition,

$$P_0(\Omega) \equiv P_0\left(\prod_{t\in I} M_t\right) = \nu_{t_1}(M_{t_1}) = 1.$$

Next I will show P_0 is a finite measure on $\mathscr{E}$. After this it is only a matter of using the Caratheodory extension theorem to get the existence of the desired probability measure P.

Claim: Suppose $\boldsymbol{E}^n$ is in $\mathscr{E}$ and suppose $\boldsymbol{E}^n \downarrow \emptyset$. Then $P_0(\boldsymbol{E}^n) \downarrow 0$.

Proof of the claim: If not, there exists a sequence such that although $\boldsymbol{E}^n \downarrow \emptyset, P_0(\boldsymbol{E}^n) \downarrow \varepsilon > 0$. Let $\boldsymbol{E}^n \in \mathscr{E}_{J_n}$. Thus it is a finite disjoint union of sets of $\mathscr{R}_{J_n}$. By regularity of the measures ν_J, which follows from Lemmas 9.4.4 and 9.4.5, there exists $\boldsymbol{K}_{J_n} \subseteq \boldsymbol{E}^n$ such that

$$\nu_{J_n}(\pi_{J_n}(\boldsymbol{K}_{J_n})) + \frac{\varepsilon}{2^{n+2}} > \nu_{J_n}(\pi_{J_n}(\boldsymbol{E}^n))$$

Thus

$$\begin{aligned} P_0(\boldsymbol{K}_{J_n}) + \frac{\varepsilon}{2^{n+2}} &\equiv \nu_{J_n}(\pi_{J_n}(\boldsymbol{K}_{J_n})) + \frac{\varepsilon}{2^{n+2}} \\ &> \nu_{J_n}(\pi_{J_n}(\boldsymbol{E}^n)) \equiv P_0(\boldsymbol{E}^n) \end{aligned}$$

The interesting thing about these $\boldsymbol{K}_{J_n}$ is: they have the finite intersection property. Here is why.

$$\begin{aligned} \varepsilon &\leq P_0\left(\cap_{k=1}^m \boldsymbol{K}_{J_k}\right) + P_0\left(\boldsymbol{E}^m \setminus \cap_{k=1}^m \boldsymbol{K}_{J_k}\right) \\ &\leq P_0\left(\cap_{k=1}^m \boldsymbol{K}_{J_k}\right) + P_0\left(\cup_{k=1}^m \boldsymbol{E}^k \setminus \boldsymbol{K}_{J_k}\right) \\ &< P_0\left(\cap_{k=1}^m \boldsymbol{K}_{J_k}\right) + \sum_{k=1}^{\infty} \frac{\varepsilon}{2^{k+2}} < P_0\left(\cap_{k=1}^m \boldsymbol{K}_{J_k}\right) + \varepsilon/2, \end{aligned}$$

and so $P_0\left(\cap_{k=1}^m \boldsymbol{K}_{J_k}\right) > \varepsilon/2$. In considering all the $\boldsymbol{E}^n$, there are countably many entries in the product space which have something other than M_t in them. Say these are $\{t_1, t_2, \cdots\}$. Let p_{t_i} be a point which is in the intersection of the t_i components of the sets $\boldsymbol{K}_{J_n}$. The compact sets in the t_i position must have the finite intersection property also because if not, the sets $\boldsymbol{K}_{J_n}$ can't have it. Thus there is such a point. As to the other positions, use the axiom of choice to pick something in each of these. Thus the intersection of these $\boldsymbol{K}_{J_n}$ contains a point which is contrary to $\boldsymbol{E}^n \downarrow \emptyset$ because these sets are contained in the $\boldsymbol{E}^n$.

With the claim, it follows P_0 is a measure on $\mathscr{E}$. Here is why: If $\boldsymbol{E} = \cup_{k=1}^{\infty} \boldsymbol{E}^k$ where $\boldsymbol{E}, \boldsymbol{E}^k \in \mathscr{E}$, then $(\boldsymbol{E} \setminus \cup_{k=1}^n \boldsymbol{E}_k) \downarrow \emptyset$ and so

$$P_0(\cup_{k=1}^n \boldsymbol{E}_k) \to P_0(\boldsymbol{E}).$$

Hence if the $\boldsymbol{E}_k$ are disjoint, $P_0\left(\cup_{k=1}^n \boldsymbol{E}_k\right) = \sum_{k=1}^n P_0(\boldsymbol{E}_k) \to P_0(\boldsymbol{E})$. Thus for disjoint $\boldsymbol{E}_k$ having $\cup_k \boldsymbol{E}_k = \boldsymbol{E} \in \mathscr{E}$,

$$P_0(\cup_{k=1}^{\infty} \boldsymbol{E}_k) = \sum_{k=1}^{\infty} P_0(\boldsymbol{E}_k).$$

Now to conclude the proof, apply the Caratheodory extension theorem to obtain P a probability measure which extends P_0 to a σ algebra which contains $\sigma(\mathscr{E})$ the sigma algebra generated by $\mathscr{E}$ with $P = P_0$ on $\mathscr{E}$. Thus for $\boldsymbol{E}_J \in \mathscr{E}$, $P(\boldsymbol{E}_J) = P_0(\boldsymbol{E}_J) = \nu_J(P_J \boldsymbol{E}_j)$.

Next, let $(\prod_{t\in I} M_t, \mathscr{F}, P)$ be the probability space and for $\boldsymbol{x} \in \prod_{t\in I} M_t$ let $X_t(\boldsymbol{x}) = x_t$, the t^{th} entry of $\boldsymbol{x}$. It follows X_t is measurable (also continuous) because if U is open in M_t, then $X_t^{-1}(U)$ has a U in the t^{th} slot and M_s everywhere else for $s \neq t$. Thus inverse images of open sets are measurable. Also, letting J be a finite subset of I and for $J = (t_1, \cdots, t_n)$, and $F_{t_1}, \cdots, F_{t_n}$ Borel sets in $M_{t_1} \cdots M_{t_n}$ respectively, it follows $\boldsymbol{F}_J$, where $\boldsymbol{F}_J$ has F_{t_i} in the t_i^{th} entry, is in $\mathscr{E}$ and therefore,

$$P([X_{t_1} \in F_{t_1}] \cap [X_{t_2} \in F_{t_2}] \cap \cdots \cap [X_{t_n} \in F_{t_n}]) =$$

$$P([(X_{t_1}, X_{t_2}, \cdots, X_{t_n}) \in F_{t_1} \times \cdots \times F_{t_n}]) = P(\boldsymbol{F}_J) = P_0(\boldsymbol{F}_J)$$

$$= \nu_{t_1 \cdots t_n}(F_{t_1} \times \cdots \times F_{t_n})$$

Finally consider the claim about the integrals. Suppose $f(x_{t_1}, \cdots, x_{t_n}) = \mathscr{X}_F$ where F is a Borel set of $\prod_{t\in J} M_t$ where $J = (t_1, \cdots, t_n)$. To begin with suppose

$$F = F_{t_1} \times \cdots \times F_{t_n} \tag{20.4}$$

where each F_{t_j} is in $\mathscr{B}(M_{t_j})$. Then

$$\int_{M_{t_1} \times \cdots \times M_{t_n}} \mathscr{X}_F(x_{t_1}, \cdots, x_{t_n})\, d\nu_{t_1 \cdots t_n} = \nu_{t_1 \cdots t_n}(F_{t_1} \times \cdots \times F_{t_n})$$

$$\begin{aligned} &= P\left(\prod_{t\in I} F_t\right) = \int_\Omega \mathscr{X}_{\prod_{t\in I} F_t}(\boldsymbol{x})\, dP \\ &= \int_\Omega \mathscr{X}_F(x_{t_1}, \cdots, x_{t_n})\, dP \end{aligned} \tag{20.5}$$

where $F_t = M_t$ if $t \notin J$. Let $\mathscr{K}$ denote sets F of the sort in 20.4. It is clearly a π system. Now let $\mathscr{G}$ denote those sets F in $\mathscr{B}(\prod_{t\in J} M_t)$ such that 20.5 holds. Thus $\mathscr{G} \supseteq \mathscr{K}$. It is clear that $\mathscr{G}$ is closed with respect to countable disjoint unions and complements. Hence $\mathscr{G} \supseteq \sigma(\mathscr{K})$ but $\sigma(\mathscr{K}) = \mathscr{B}(\prod_{t\in J} M_t)$ because every open set in $\prod_{t\in J} M_t$ is the countable union of rectangles like 20.4 in which each F_{t_i} is open. Therefore, 20.5 holds for every $F \in \mathscr{B}(\prod_{t\in J} M_t)$.

Passing to simple functions and then using the monotone convergence theorem yields the final claim of the theorem. ■

As a special case, you can obtain a version of product measure for possibly infinitely many factors. Suppose in the context of the above theorem that ν_t is a probability measure defined on the Borel sets of $M_t \equiv \mathbb{R}^{n_t}$ for n_t a positive integer, and let the measures, $\nu_{t_1 \cdots t_n}$ be defined on the Borel sets of $\prod_{i=1}^n M_{t_i}$ by

$$\nu_{t_1 \cdots t_n}(\boldsymbol{E}) \equiv \overbrace{(\nu_{t_1} \times \cdots \times \nu_{t_n})}^{\text{product measure}}(\boldsymbol{E}).$$

Then these measures satisfy the necessary consistency condition and so the Kolmogorov extension theorem given above can be applied to obtain a measure P defined on a measure space $(\prod_{t\in I} M_t, \mathscr{F})$ and measurable functions $X_s : \prod_{t\in I} M_t \to M_s$ such that for F_{t_i} a Borel set in M_{t_i},

$$P\left((X_{t_1}, \cdots, X_{t_n}) \in \prod_{i=1}^n F_{t_i}\right) = \nu_{t_1 \cdots t_n}(F_{t_1} \times \cdots \times F_{t_n})$$

$$= \nu_{t_1}(F_{t_1}) \cdots \nu_{t_n}(F_{t_n}). \tag{20.6}$$

In particular, $P(X_t \in F_t) = \nu_t(F_t)$. Then P in the resulting probability space,

$$\left(\prod_{t\in I} M_t, \mathscr{F}, P\right)$$

will be denoted as $\prod_{t\in I} \nu_t$. This proves the following theorem which describes an infinite product measure.

Theorem 20.3.4 *Let M_t for $t \in I$ be given as in Theorem 20.3.3 and let ν_t be a Borel probability measure defined on the Borel sets of M_t. Then there exists a measure P and a σ algebra $\mathscr{F} = \sigma(\mathscr{E})$ where $\mathscr{E}$ is given in the Notation 20.3.1 such that $(\prod_t M_t, \mathscr{F}, P)$ is a probability space satisfying 20.6 whenever each F_{t_i} is a Borel set of M_{t_i}. This probability measure could be denoted as $\prod_t \nu_t$.*

20.4 Exercises

1. Suppose X and Y are metric spaces having compact closed balls. Show $(X \times Y, d_{X\times Y})$ is also a metric space which has the closures of balls compact. Here

$$d_{X\times Y}((x_1,y_1),(x_2,y_2)) \equiv \max(d(x_1,x_2), d(y_1,y_2)).$$

   Let

$$\mathscr{A} \equiv \{E \times F : E \text{ is a Borel set in } X, F \text{ is a Borel set in } Y\}.$$

   Show $\sigma(\mathscr{A})$, the smallest σ algebra containing $\mathscr{A}$ contains the Borel sets. **Hint:** Show every open set in a metric space which has closed balls compact can be obtained as a countable union of compact sets. Next show this implies every open set can be obtained as a countable union of open sets of the form $U \times V$ where U is open in X and V is open in Y.

2. Suppose $(\Omega, \mathscr{S}, \mu)$ is a measure space which may not be complete. Could you obtain a complete measure space, $(\Omega, \overline{\mathscr{S}}, \mu_1)$ by simply letting $\overline{\mathscr{S}}$ consist of all sets of the form E where there exists $F \in \mathscr{S}$ such that $(F \setminus E) \cup (E \setminus F) \subseteq N$ for some $N \in \mathscr{S}$ which has measure zero and then let $\mu(E) = \mu_1(F)$? Explain.

3. Let $(\Omega, \mathscr{S}, \mu)$ measure space and let $f : \Omega \to [0,\infty)$ be measurable. Define $A \equiv \{(x,y) : y < f(x)\}$. Show that $\int f d\mu = \int\int \mathscr{X}_A(x,y)\, d\mu dm$ Next show that A is product measurable in the sense that A_x is m measurable and A_y is μ measurable. Here $A_x \equiv \{y : (x,y) \in A\}$ and A_y similar. Next show that you can interchange the order of integration. **Hint:** First suppose f is a nonnegative simple function.

4. For f a nonnegative measurable function, it was shown that

$$\int f d\mu = \int \mu([f > t])\, dt.$$

   Would it work the same if you used $\int \mu([f \geq t])\, dt$? Explain.

5. Let $(\Omega, \mathscr{F}, \mu)$ be a finite measure space and suppose $\{f_n\}$ is a sequence of nonnegative functions which satisfy $f_n(\omega) \leq C$ independent of n, ω. Suppose also this sequence converges to 0 in measure. That is, for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \mu([f_n \geq \varepsilon]) = 0$$

Show that then $\lim_{n\to\infty} \int_\Omega f_n(\omega)\, d\mu = 0$.

6. Explain why for each $t > 0, x \to e^{-tx}$ is a function in $L^1(\mathbb{R})$ and

$$\int_0^\infty e^{-tx} dx = \frac{1}{t}.$$

Thus $\int_0^R \frac{\sin(t)}{t} dt = \int_0^R \int_0^\infty \sin(t) e^{-tx} dx dt$ Now explain why you can change the order of integration in the above iterated integral. Then compute what you get. Next pass to a limit as $R \to \infty$ and show $\int_0^\infty \frac{\sin(t)}{t} dt = \frac{1}{2}\pi$

7. Let $f(y) = g(y) = |y|^{-1/2}$ if $y \in (-1,0) \cup (0,1)$ and

$$f(y) = g(y) = 0$$

if $y \notin (-1,0) \cup (0,1)$. For which values of x does it make sense to write the integral

$$\int_{\mathbb{R}} f(x-y)\, g(y)\, dy?$$

8. Let E_i be a Borel set in $\mathbb{R}$. Show that $\prod_{i=1}^n E_i$ is a Borel set in $\mathbb{R}^n$.

9. Let $\{a_n\}$ be an increasing sequence of numbers in $(0,1)$ which converges to 1. Let g_n be a nonnegative function which equals zero outside (a_n, a_{n+1}) such that $\int g_n dx = 1$. Now for $(x,y) \in [0,1) \times [0,1)$ define

$$f(x,y) \equiv \sum_{k=1}^\infty g_n(y)\,(g_n(x) - g_{n+1}(x)).$$

Explain why this is actually a finite sum for each such (x,y) so there are no convergence questions in the infinite sum. Explain why f is a continuous function on $[0,1) \times [0,1)$. You can extend f to equal zero off $[0,1) \times [0,1)$ if you like. Show the iterated integrals exist but are not equal. In fact, show

$$\int_0^1 \int_0^1 f(x,y)\, dy dx = 1 \neq 0 = \int_0^1 \int_0^1 f(x,y)\, dx dy.$$

Does this example contradict the Fubini theorem, Corollary 10.14.10 on Page 314? Explain why or why not.

10. Let $f : [a,b] \to \mathbb{R}$ be Rieman integrable. Thus f is a bounded function and by Darboux's theorem, there exists a unique number between all the upper sums and lower sums of f, this number being the Riemann integral. Show that f is Lebesgue measurable and $\int_a^b f(x)\, dx = \int_{[a,b]} f dm$ where the second integral in the above is the Lebesgue integral taken with respect to one dimensional Lebesgue measure and the first is the ordinary Riemann integral.

11. Let $(\Omega, \mathscr{F}, \mu)$ be a σ finite measure space and let $f : \Omega \to [0,\infty)$ be measurable. Also let $\phi : [0,\infty) \to \mathbb{R}$ be increasing with $\phi(0) = 0$ and ϕ a C^1 function. Show that

$$\int_\Omega \phi \circ f d\mu = \int_0^\infty \phi'(t)\, \mu([f > t])\, dt.$$

Hint: This can be done using the following steps. Let $t_i^n = i2^{-n}$. Show that

$$\mathscr{X}_{[f>t]}(\omega) = \lim_{n\to\infty} \sum_{i=0}^{\infty} \mathscr{X}_{\left[f > t_{i+1}^n\right]}(\omega)\, \mathscr{X}_{[t_i^n, t_{i+1}^n)}(t)$$

Now this is a countable sum of $\mathscr{F} \times \mathscr{B}([0,\infty))$ measurable functions and so it follows that $(t,\omega) \to \mathscr{X}_{[f>t]}(\omega)$ is $\mathscr{F} \times \mathscr{B}([0,\infty))$ measurable. Consequently, so is $\mathscr{X}_{[f>t]}(\omega)\phi(t)$. Note that it is important in the argument to have $f > t$. Now observe

$$\int_\Omega \phi \circ f d\mu = \int_\Omega \int_0^{f(\omega)} \phi'(t)\, dt d\mu = \int_\Omega \int_0^\infty \mathscr{X}_{[f>t]}(\omega)\, \phi'(t)\, dt d\mu$$

Use Fubini's theorem. For your information, this does not require the measure space to be σ finite. You can use a different argument which ties in to the first definition of the Lebesgue integral. The function $t \to \mu([f > t])$ is called the distribution function.

12. Give a different proof of the above as follows. First suppose f is a simple function, $f(\omega) = \sum_{k=1}^n a_k \mathscr{X}_{E_k}(\omega)$ where the a_k are strictly increasing, $\phi(a_0) = a_0 \equiv 0$. Then explain carefully the steps to the following argument.

$$\begin{aligned}
\int_\Omega \phi \circ f d\mu &= \sum_{i=1}^n \int_{\phi(a_{i-1})}^{\phi(a_i)} \mu([\phi \circ f > t])\, dt = \sum_{i=1}^n \int_{\phi(a_{i-1})}^{\phi(a_i)} \sum_{k=i}^n \mu(E_k)\, dt \\
&= \sum_{i=1}^n \sum_{k=i}^n \mu(E_k) \int_{a_{i-1}}^{a_i} \phi'(t)\, dt = \sum_{i=1}^n \int_{a_{i-1}}^{a_i} \phi'(t) \sum_{k=i}^n \mu(E_k)\, dt \\
&= \sum_{i=1}^n \int_{a_{i-1}}^{a_i} \phi'(t)\, \mu([f > t])\, dt = \int_0^\infty \phi'(t)\, \mu([f > t])\, dt
\end{aligned}$$

Note that this did not require the measure space to be σ finite and comes directly from the definition of the integral.

13. Give another argument for the above result as follows.

$$\int \phi \circ f d\mu = \int_0^\infty \mu([\phi \circ f > t])\, dt = \int_0^\infty \mu\left(\left[f > \phi^{-1}(t)\right]\right) dt$$

and now change the variable in the last integral, letting $\phi(s) = t$. Justify the easy manipulations.

Chapter 21

Banach Spaces

21.1 Theorems Based on Baire Category

Some examples of Banach spaces that have been discussed up to now are $\mathbb{R}^n, \mathbb{C}^n$, and $L^p(\Omega)$. Theorems about general Banach spaces are proved in this chapter. The main theorems to be presented here are the uniform boundedness theorem, the open mapping theorem, the closed graph theorem, and the Hahn Banach Theorem. The first three of these theorems come from the Baire category theorem which is about to be presented. They are topological in nature. The Hahn Banach theorem has nothing to do with topology. Banach spaces are all normed linear spaces and as such, they are all metric spaces because a normed linear space may be considered as a metric space with $d(x,y) \equiv ||x-y||$. You can check that this satisfies all the axioms of a metric. As usual, if every Cauchy sequence converges, the metric space is called complete.

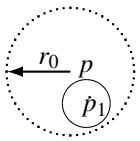
Definition 21.1.1 *A complete normed linear space is called a Banach space.*

21.1.1 Baire Category Theorem

The following remarkable result is called the Baire category theorem. To get an idea of its meaning, imagine you draw a line in the plane. The complement of this line is an open set and is dense because every point, even those on the line, are limit points of this open set. Now draw another line. The complement of the two lines is still open and dense. Keep drawing lines and looking at the complements of the union of these lines. You always have an open set which is dense. Now what if there were countably many lines? The Baire category theorem implies the complement of the union of these lines is dense. In particular it is nonempty. Thus you cannot write the plane as a countable union of lines. This is a rather rough description of this very important theorem. The precise statement and proof follow.

Theorem 21.1.2 *Let (X,d) be a complete metric space and let $\{U_n\}_{n=1}^{\infty}$ be a sequence of open subsets of X satisfying $\overline{U_n} = X$ (U_n is dense). Then $D \equiv \cap_{n=1}^{\infty} U_n$ is a dense subset of X.*

Proof: Let $p \in X$ and let $r_0 > 0$. I need to show $D \cap B(p,r_0) \neq \emptyset$. Since U_1 is dense, there exists $p_1 \in U_1 \cap B(p,r_0)$, an open set. Let $p_1 \in B(p_1,r_1) \subseteq \overline{B(p_1,r_1)} \subseteq U_1 \cap B(p,r_0)$ and $r_1 < 2^{-1}$. This is possible because $U_1 \cap B(p,r_0)$ is an open set and so there exists r_1 such that $B(p_1,2r_1) \subseteq U_1 \cap B(p,r_0)$. But $B(p_1,r_1) \subseteq \overline{B(p_1,r_1)} \subseteq B(p_1,2r_1)$ because $\overline{B(p_1,r_1)} \subseteq \{x \in X : d(x,p_1) \leq r_1\} \leq B(p_1,2r_1)$. Indeed, $\{x \in X : d(x,p_1) \leq r_1\}$ is a closed set containing $B(p_1,r)$ so it contains $\overline{B(p_1,r_1)}$.



There exists $p_2 \in U_2 \cap B(p_1,r_1)$ because U_2 is dense. Let

$$p_2 \in B(p_2,r_2) \subseteq \overline{B(p_2,r_2)} \subseteq U_2 \cap B(p_1,r_1) \subseteq U_1 \cap U_2 \cap B(p,r_0).$$

and let $r_2 < 2^{-2}$. Continue in this way. Thus $r_n < 2^{-n}$,

$$\overline{B(p_n, r_n)} \subseteq U_1 \cap U_2 \cap \ldots \cap U_n \cap B(p, r_0),$$

$$\overline{B(p_n, r_n)} \subseteq B(p_{n-1}, r_{n-1}).$$

The sequence, $\{p_n\}$ is a Cauchy sequence because all terms of $\{p_k\}$ for $k \geq n$ are contained in $B(p_n, r_n)$, a set whose diameter is no larger than 2^{-n}. Since X is complete, there exists p_∞ such that $\lim_{n\to\infty} p_n = p_\infty$. Since all but finitely many terms of $\{p_n\}$ are in $\overline{B(p_m, r_m)}$, it follows that $p_\infty \in \overline{B(p_m, r_m)}$ for each m. Therefore, $p_\infty \in \cap_{m=1}^{\infty} \overline{B(p_m, r_m)} \subseteq \cap_{i=1}^{\infty} U_i \cap B(p, r_0)$. ■

The following corollary is also called the Baire category theorem. It involves a countable union of closed sets rather than a countable intersection of open sets.

Corollary 21.1.3 *Let X be a complete metric space and suppose $X = \cup_{i=1}^{\infty} F_i$ where each F_i is a closed set. Then for some i, interior $F_i \neq \emptyset$.*

Proof: If all F_i has empty interior, then F_i^C would be a dense open set. Therefore, from Theorem 21.1.2, $\emptyset = (\cup_{i=1}^{\infty} F_i)^C = \cap_{i=1}^{\infty} F_i^C \neq \emptyset$. ■

The set D of Theorem 21.1.2 is called a G_δ set because it is the countable intersection of open sets. Thus D is a dense G_δ set.

Recall that a norm satisfies:

a.) $\|x\| \geq 0$, $\|x\| = 0$ if and only if $x = 0$.

b.) $\|x+y\| \leq \|x\| + \|y\|$.

c.) $\|cx\| = |c| \, \|x\|$ if c is a scalar and $x \in X$.

From Theorem 3.6.2, continuity means that if $\lim_{n\to\infty} x_n = x$, then

$$\lim_{n\to\infty} f(x_n) = f(x).$$

Theorem 21.1.4 *Let X and Y be two normed linear spaces and let $L : X \to Y$ be linear ($L(ax+by) = aL(x) + bL(y)$ for a, b scalars and $x, y \in X$). The following are equivalent*

a.) L is continuous at 0

b.) L is continuous

c.) There exists $K > 0$ such that $\|Lx\|_Y \leq K \|x\|_X$ for all $x \in X$ (L is bounded).

Proof: a.)⇒b.) Let $x_n \to x$. It is necessary to show that $Lx_n \to Lx$. But $(x_n - x) \to 0$ and so from continuity at 0, it follows $L(x_n - x) = Lx_n - Lx \to 0$ so $Lx_n \to Lx$. This shows a.) implies b.).

b.)⇒c.) Since L is continuous, L is continuous at 0. Hence $\|Lx\|_Y < 1$ whenever $\|x\|_X \leq \delta$ for some δ. Therefore, suppressing the subscript on the $\| \ \|$, $\|L\left(\frac{\delta x}{\|x\|}\right)\| \leq 1$. Hence $\|Lx\| \leq \frac{1}{\delta}\|x\|$.

c.)⇒a.) follows from the inequality given in c.). ■

Definition 21.1.5 *Let $L : X \to Y$ be linear and continuous where X and Y are normed linear spaces. Denote the set of all such continuous linear maps by $\mathscr{L}(X,Y)$ and define*

$$\|L\| = \sup\{\|Lx\| : \|x\| \leq 1\}. \tag{21.1}$$

This is called the operator norm.

Note that from Theorem 21.1.4 $||L||$ is well defined because of part c.) of that Theorem.

The next lemma follows immediately from the definition of the norm and the assumption that L is linear.

Lemma 21.1.6 *With $\|L\|$ defined in 21.1, $\mathscr{L}(X,Y)$ is a normed linear space. Also $\|Lx\| \leq \|L\| \|x\|$.*

Proof: Let $x \neq 0$ then $x/\|x\|$ has norm equal to 1 and so $\left\|L\left(\frac{x}{\|x\|}\right)\right\| \leq \|L\|$. Therefore, multiplying both sides by $\|x\|$, $\|Lx\| \leq \|L\| \|x\|$. This is obviously a linear space. It remains to verify the operator norm really is a norm. First of all, if $\|L\| = 0$, then $Lx = 0$ for all $\|x\| \leq 1$. It follows that for any $x \neq 0, 0 = L\left(\frac{x}{\|x\|}\right)$ and so $Lx = 0$. Therefore, $L = 0$. Also, if c is a scalar,

$$\|cL\| = \sup_{\|x\|\leq 1} \|cL(x)\| = |c| \sup_{\|x\|\leq 1} \|Lx\| = |c| \|L\|.$$

It remains to verify the triangle inequality. Let $L, M \in \mathscr{L}(X,Y)$.

$$\begin{aligned} \|L+M\| &\equiv \sup_{\|x\|\leq 1} \|(L+M)(x)\| \leq \sup_{\|x\|\leq 1} (\|Lx\| + \|Mx\|) \\ &\leq \sup_{\|x\|\leq 1} \|Lx\| + \sup_{\|x\|\leq 1} \|Mx\| = \|L\| + \|M\|. \end{aligned}$$

This shows the operator norm is really a norm as hoped. ■

For example, consider the space of linear transformations defined on $\mathbb{R}^n$ having values in $\mathbb{R}^m$. The fact the transformation is linear automatically imparts continuity to it. You should give a proof of this fact. Recall that every such linear transformation can be realized in terms of matrix multiplication.

Thus, in finite dimensions the algebraic condition that an operator is linear is sufficient to imply the topological condition that the operator is continuous. The situation is not so simple in infinite dimensional spaces such as $C(X;\mathbb{R}^n)$. This explains the imposition of the topological condition of continuity as a criterion for membership in $\mathscr{L}(X,Y)$ in addition to the algebraic condition of linearity. Here is an example which shows that this extra assumption cannot be eliminated.

Example 21.1.7 *Let V denote all linear combinations of functions of the form $e^{-\alpha x^2}$ for $\alpha > 0$. Thus a typical element of V is an expression of the form*

$$\sum_{k=1}^{n} \beta_k e^{-\alpha_k x^2}, \alpha_k > 0.$$

Let $L : V \to \mathbb{C}$ be given by $Lf \equiv \int_{\mathbb{R}} f(x)\,dx$. For a norm on V, $\|f\| \equiv \max\{|f(x)| : x \in \mathbb{R}\}$. Of course V is not complete, but it is a normed linear space and you could consider its completion if desired, in terms of equivalence classes of Cauchy sequences, similar to the construction of $\mathbb{R}$ from $\mathbb{Q}$. Recall that $\int_{-\infty}^{\infty} e^{-x^2} dx = \int_{-\infty}^{\infty} \frac{1}{n} e^{-\left(x^2/n^2\right)} = \sqrt{\pi}$ where here $n \in \mathbb{N}$. Consider the sequence of functions $f_n(x) \equiv \frac{1}{n} e^{-\left(x^2/n^2\right)}$. Its maximum value is $1/n$ and so $\|f_n\| \to 0$ but Lf_n fails to converge to 0. Thus L is not continuous although it is linear.

Theorem 21.1.8 *If Y is a Banach space, then $\mathscr{L}(X,Y)$ is also a Banach space.*

Proof: Let $\{L_n\}$ be a Cauchy sequence in $\mathscr{L}(X,Y)$ and let $x \in X$.

$$\|L_n x - L_m x\| \leq \|x\| \, \|L_n - L_m\|.$$

Thus $\{L_n x\}$ is a Cauchy sequence. Let $Lx = \lim_{n\to\infty} L_n x$. Then, clearly, L is linear because if x_1, x_2 are in X, and a, b are scalars, then

$$L(ax_1 + bx_2) = \lim_{n\to\infty} L_n(ax_1 + bx_2) = \lim_{n\to\infty} (aL_n x_1 + bL_n x_2) = aLx_1 + bLx_2.$$

Also L is continuous. To see this, note that $\{\|L_n\|\}$ is a Cauchy sequence of real numbers because $|\|L_n\| - \|L_m\|| \leq \|L_n - L_m\|$. Hence there exists $K > \sup\{\|L_n\| : n \in \mathbb{N}\}$. Thus, if $x \in X$, $\|Lx\| = \lim_{n\to\infty} \|L_n x\| \leq K \|x\|$.■

21.1.2 Uniform Boundedness Theorem

The next big result is sometimes called the Uniform Boundedness theorem, or the Banach-Steinhaus theorem. This is a very surprising theorem which implies that for a collection of bounded linear operators, if they are bounded pointwise, then they are also bounded uniformly. As an example of a situation in which pointwise bounded does not imply uniformly bounded, consider the functions $f_\alpha(x) \equiv \mathscr{X}_{(\alpha,1)}(x) x^{-1}$ for $\alpha \in (0,1)$. Clearly each function is bounded and the collection of functions is bounded at each point of $(0,1)$, but there is no bound for all these functions taken together.

Theorem 21.1.9 *Let X be a Banach space and let Y be a normed linear space. Let $\{L_\alpha\}_{\alpha\in\Lambda}$ be a collection of elements of $\mathscr{L}(X,Y)$. Then one of the following happens.*

a.) $\sup\{\|L_\alpha\| : \alpha \in \Lambda\} < \infty$

b.) There exists a dense G_δ set D, such that for all $x \in D$,

$$\sup\{\|L_\alpha x\| \ \alpha \in \Lambda\} = \infty.$$

Proof: For each $n \in \mathbb{N}$, define $U_n = \{x \in X : \sup\{\|L_\alpha x\| : \alpha \in \Lambda\} > n\}$. Then U_n is an open set because if $x \in U_n$, then there exists $\alpha \in \Lambda$ such that $\|L_\alpha x\| > n$. But then, since L_α is continuous, this situation persists for all y sufficiently close to x, say for all $y \in B(x,\delta)$. Then $B(x,\delta) \subseteq U_n$ which shows U_n is open.

Case b.) is obtained from Theorem 21.1.2 if each U_n is dense.

The other case is that for some n, U_n is not dense. If this occurs, there exists x_0 and $r > 0$ such that for all $B(x_0,r) \subseteq U_n^C$ so for all $x \in B(x_0,r)$, $\|L_\alpha x\| \leq n$ for all α. Now if $y \in B(0,r)$, $x_0 + y \in B(x_0,r)$. Consequently, for all such y, $\|L_\alpha(x_0+y)\| \leq n$. This implies that for all $\alpha \in \Lambda$ and $\|y\| < r$, $\|L_\alpha y\| \leq n + \|L_\alpha(x_0)\| \leq 2n$. Therefore, if $\|y\| \leq 1$, $\left\|\frac{r}{2}y\right\| < r$ and so for all α, $\left\|L_\alpha\left(\frac{r}{2}y\right)\right\| \leq 2n$. Now multiplying by $r/2$ it follows that whenever $\|y\| \leq 1$, $\|L_\alpha(y)\| \leq 4n/r$. Hence case a.) holds. ■

21.1.3 Open Mapping Theorem

Another remarkable theorem which depends on the Baire category theorem is the open mapping theorem. Unlike Theorem 21.1.9 it requires both X and Y to be Banach spaces.

Theorem 21.1.10 *Let X and Y be Banach spaces, let $L \in \mathscr{L}(X,Y)$, and suppose L is onto. Then L maps open sets onto open sets.*

To aid in the proof, here is a lemma.

Lemma 21.1.11 *Let a and b be positive constants and suppose $B(0,a) \subseteq \overline{L(B(0,b))}$. Then $\overline{L(B(0,b))} \subseteq L(B(0,2b))$.*

Proof: Let $z \in \overline{L(B(0,b))}$. Then let $x_1 \in B(0,b)$ with $\|z - Lx_1\| < \frac{a}{2^1}$. Then it follows that on multiplying by 2, $\|2z - 2Lx_1\| < a$ and so there is $x_2 \in B(0,b)$ with

$$\|(2z - 2Lx_1) - Lx_2\| < \frac{a}{2}$$

which implies $\left\|z - \left(Lx_1 + \frac{Lx_2}{2}\right)\right\| < \frac{a}{2^2}$. If $x_i \in B(0,b)$ and $\left\|z - \sum_{i=0}^{n} \frac{Lx_i}{2^i}\right\| < \frac{a}{2^n}$, then

$$\left\|2^n \left(z - \sum_{i=0}^{n} \frac{Lx_i}{2^i}\right)\right\| < a$$

Continuing this way, there is $x_{n+1} \in B(0,b)$ such that

$$\left\|2^n \left(z - \sum_{i=0}^{n} \frac{Lx_i}{2^i}\right) - L(x_{n+1})\right\| < \frac{a}{2}$$

and so $\left\|z - \sum_{i=0}^{n+1} \frac{Lx_i}{2^i}\right\| < \frac{a}{2^{n+1}}$. Let $x \equiv \sum_{i=0}^{\infty} \frac{x_i}{2^i}$. Then from the triangle inequality, $\|x\| < \sum_{i=0}^{\infty} \frac{b}{2^i} = 2b$ and by continuity of L,

$$\|z - Lx\| = \lim_{n\to\infty} \left\|z - L\left(\sum_{i=0}^{n} \frac{x_i}{2^i}\right)\right\| \leq \lim_{n\to\infty} \frac{a}{2^n} = 0 \blacksquare$$

Proof of Theorem 21.1.10: $Y = \cup_{n=1}^{\infty} \overline{L(B(0,n))}$. By Corollary 21.1.3, $\overline{L(B(0,n_0))}$ has nonempty interior for some n_0. Thus $B(y,r) \subseteq \overline{L(B(0,n_0))}$ for some y and some $r > 0$. Since L is linear $B(-y,r) \subseteq \overline{L(B(0,n_0))}$ also. Here is why. If $z \in B(-y,r)$, then $-z \in B(y,r)$ and so there exists $x_n \in B(0,n_0)$ such that $Lx_n \to -z$. Therefore, $L(-x_n) \to z$ and $-x_n \in B(0,n_0)$ also. Therefore $z \in \overline{L(B(0,n_0))}$. Then it follows that

$$\begin{array}{rcl} B(0,r) & \subseteq & B(y,r) + B(-y,r) \equiv \{y_1 + y_2 : y_1 \in B(y,r) \text{ and } y_2 \in B(-y,r)\} \\ & \subseteq & \overline{L(B(0,2n_0))} \end{array}$$

The reason for the last inclusion is that from the above, if $y_1 \in B(y,r)$ and $y_2 \in B(-y,r)$, there exists $x_n, z_n \in B(0,n_0)$ such that $Lx_n \to y_1$, $Lz_n \to y_2$. Therefore, $\|x_n + z_n\| \leq 2n_0$ and so $(y_1 + y_2) \in \overline{L(B(0,2n_0))}$.

By Lemma 21.1.11, $\overline{L(B(0,2n_0))} \subseteq L(B(0,4n_0))$ so $B(0,r) \subseteq L(B(0,4n_0))$. Letting $a = r(4n_0)^{-1}$, it follows, since L is linear, that $B(0,a) \subseteq L(B(0,1))$. It follows since L is linear,

$$L(B(0,r)) \supseteq B(0,ar). \tag{21.2}$$

Now let U be open in X and let $x + B(0,r) = B(x,r) \subseteq U$. Using 21.2,

$$L(U) \supseteq L(x + B(0,r))$$

$$= Lx + L(B(0,r)) \supseteq Lx + B(0,ar) = B(Lx,ar).$$

Hence $Lx \in B(Lx, ar) \subseteq L(U)$ which shows that every point, $Lx \in LU$, is an interior point of LU and so LU is open. ■

This theorem is surprising because it implies that if $|\cdot|$ and $\|\cdot\|$ are two norms with respect to which a vector space X is a Banach space such that $|\cdot| \leq K\|\cdot\|$, then there exists a constant k, such that $\|\cdot\| \leq k|\cdot|$. This can be useful because sometimes it is not clear how to compute k when all that is needed is its existence. To see the open mapping theorem implies this, consider the identity map $\mathrm{id}x = x$. Then $\mathrm{id} : (X, \|\cdot\|) \to (X, |\cdot|)$ is continuous and onto. Hence id is an open map which implies id^{-1} is continuous. Theorem 21.1.4 gives the existence of the constant k.

21.1.4 Closed Graph Theorem

Definition 21.1.12 *Let $f : D \to E$. The set of all ordered pairs of the form*

$$\{(x, f(x)) : x \in D\}$$

is called the graph of f.

Definition 21.1.13 *If X and Y are normed linear spaces, make $X \times Y$ into a normed linear space by using the norm $\|(x,y)\| = \max(\|x\|, \|y\|)$ along with componentwise addition and scalar multiplication. Thus $a(x,y) + b(z,w) \equiv (ax + bz, ay + bw)$.*

There are other ways to give a norm for $X \times Y$. For example, you could define $\|(x,y)\| = \|x\| + \|y\|$

Lemma 21.1.14 *The norm defined in Definition 21.1.13 on $X \times Y$ along with the definition of addition and scalar multiplication given there make $X \times Y$ into a normed linear space.*

Proof: The only axiom for a norm which is not obvious is the triangle inequality. Therefore, consider

$$\begin{aligned} \|(x_1,y_1) + (x_2,y_2)\| &= \|(x_1 + x_2, y_1 + y_2)\| \\ &= \max(\|x_1 + x_2\|, \|y_1 + y_2\|) \\ &\leq \max(\|x_1\| + \|x_2\|, \|y_1\| + \|y_2\|) \end{aligned}$$

Both $\|x_1\| + \|x_2\|$ and $\|y_1\| + \|y_2\|$ are no larger than

$$\max(\|x_1\|, \|y_1\|) + \max(\|x_2\|, \|y_2\|)$$

and so the above is

$$\leq \max(\|x_1\|, \|y_1\|) + \max(\|x_2\|, \|y_2\|) = \|(x_1,y_1)\| + \|(x_2,y_2)\|.$$

It is obvious $X \times Y$ is a vector space from the above definition. ■

Lemma 21.1.15 *If X and Y are Banach spaces, then $X \times Y$ with the norm and vector space operations defined in Definition 21.1.13 is also a Banach space.*

Proof: The only thing left to check is that the space is complete. But this follows from the simple observation that $\{(x_n,y_n)\}$ is a Cauchy sequence in $X\times Y$ if and only if $\{x_n\}$ and $\{y_n\}$ are Cauchy sequences in X and Y respectively. Thus if $\{(x_n,y_n)\}$ is a Cauchy sequence in $X\times Y$, it follows there exist x and y such that $x_n\to x$ and $y_n\to y$. But then from the definition of the norm, $(x_n,y_n)\to(x,y)$. ■

Lemma 21.1.16 *Every closed subspace of a Banach space is a Banach space.*

Proof: If $F\subseteq X$ where X is a Banach space and $\{x_n\}$ is a Cauchy sequence in F, then since X is complete, there exists a unique $x\in X$ such that $x_n\to x$. However this means $x\in\overline{F}=F$ since F is closed. ■

Definition 21.1.17 *Let X and Y be Banach spaces and let $D\subseteq X$ be a subspace. A linear map $L:D\to Y$ is said to be closed if its graph is a closed subspace of $X\times Y$. Equivalently, L is closed if $x_n\to x$ and $Lx_n\to y$ implies $x\in D$ and $y=Lx$.*

Note the distinction between closed and continuous. If the operator is closed the assertion that $y=Lx$ only follows if it is known that the sequence $\{Lx_n\}$ converges. In the case of a continuous operator, the convergence of $\{Lx_n\}$ follows from the assumption that $x_n\to x$. It is not always the case that a mapping which is closed is necessarily continuous. Consider the function $f(x)=\tan(x)$ if x is not an odd multiple of $\frac{\pi}{2}$ and $f(x)\equiv 0$ at every odd multiple of $\frac{\pi}{2}$. Then the graph is closed and the function is defined on $\mathbb{R}$ but it clearly fails to be continuous. Of course this function is not linear. You could also consider the map,

$$\frac{d}{dx}:\left\{y\in C^1([0,1]):y(0)=0\right\}\equiv D\to C([0,1]).$$

where the norm is the uniform norm on $C([0,1]),||y||_\infty$. If $y\in D$, then $y(x)=\int_0^x y'(t)\,dt$. Therefore, if $\frac{dy_n}{dx}\to f\in C([0,1])$ and if $y_n\to y$ in $C([0,1])$ it follows that

$$\begin{array}{ccc} y_n(x) & = & \int_0^x \frac{dy_n(t)}{dx}dt \\ \downarrow & & \downarrow \\ y(x) & = & \int_0^x f(t)\,dt \end{array}$$

and so by the fundamental theorem of calculus $f(x)=y'(x)$ and so the mapping is closed. It is obviously not continuous because it takes $y(x)$ and $y(x)+\frac{1}{n}\sin(nx)$ to two functions which are far from each other even though these two functions are very close in $C([0,1])$. Furthermore, it is not defined on the whole space, $C([0,1])$.

The next theorem, the closed graph theorem, gives conditions under which closed implies continuous.

Theorem 21.1.18 *Let X and Y be Banach spaces and suppose $L:X\to Y$ is closed and linear. Then L is continuous.*

Proof: Let G be the graph of L. $G=\{(x,Lx):x\in X\}$. By Lemma 21.1.16 it follows that G is a Banach space. Define $P:G\to X$ by $P(x,Lx)=x$. P maps the Banach space G onto the Banach space X and is continuous and linear. By the open mapping theorem, P maps open sets onto open sets. Since P is also one to one, this says that P^{-1} is continuous. Thus $\left\|P^{-1}x\right\|\le K\|x\|$. Hence

$$\|Lx\|\le\max(\|x\|,\|Lx\|)\le K\|x\|$$

By Theorem 21.1.4 on Page 558, this shows L is continuous. ■

The following corollary is quite useful. It shows how to obtain a new norm on the domain of a closed operator such that the domain with this new norm becomes a Banach space.

Corollary 21.1.19 *Let $L : D \subseteq X \to Y$ where X, Y are a Banach spaces, and L is a closed operator. Then define a new norm on D by*

$$\|x\|_D \equiv \|x\|_X + \|Lx\|_Y .$$

Then D with this new norm is a Banach space.

Proof: If $\{x_n\}$ is a Cauchy sequence in D with this new norm, it follows both $\{x_n\}$ and $\{Lx_n\}$ are Cauchy sequences and therefore, they converge. Since L is closed, $x_n \to x$ and $Lx_n \to Lx$ for some $x \in D$. Thus $\|x_n - x\|_D \to 0$. ■

21.2 Hahn Banach Theorem

The closed graph, open mapping, and uniform boundedness theorems are the three major topological theorems in functional analysis. The other major theorem is the Hahn-Banach theorem which has nothing to do with topology. Before presenting this theorem, here are some preliminaries about partially ordered sets.

21.2.1 Partially Ordered Sets

Recall Theorem 2.8.2 which is stated next for convenience.

Theorem 21.2.1 *(Hausdorff Maximal Principle) Let $\mathscr{F}$ be a nonempty partially ordered set. Then there exists a maximal chain.*

21.2.2 Gauge Functions and Hahn Banach Theorem

Definition 21.2.2 *Let X be a real vector space $\rho : X \to \mathbb{R}$ is called a gauge function if*

$$\begin{aligned} &\rho(x+y) \leq \rho(x) + \rho(y), \\ &\rho(ax) = a\rho(x) \text{ if } a \geq 0. \end{aligned} \tag{21.3}$$

Suppose M is a subspace of X and $z \notin M$. Suppose also that f is a linear real-valued function having the property that $f(x) \leq \rho(x)$ for all $x \in M$. Consider the problem of extending f to $M \oplus \mathbb{R}z$ such that if F is the extended function, $F(y) \leq \rho(y)$ for all $y \in M \oplus \mathbb{R}z$ and F is linear. Since F is to be linear, it suffices to determine how to define $F(z)$. Letting $a > 0$, it is required to define $F(z)$ such that the following hold for all $x, y \in M$.

$$\begin{aligned} &\overbrace{F(x)}^{f(x)} + aF(z) = F(x+az) \leq \rho(x+az), \\ &\overbrace{F(y)}^{f(y)} - aF(z) = F(y-az) \leq \rho(y-az). \end{aligned} \tag{21.4}$$

Note that something in $M \oplus \mathbb{R}z$ is of the form $x+az$ or $x-az$ for $a \geq 0$ and this includes all possibilities. Now multiplying by a^{-1} for $a > 0$, the above holds if and only if for all $x, y \in M$,

$$F(y) - F(z) \leq \rho(y-z), F(x) + F(z) \leq \rho(x+z)$$

Thus we need to choose $F(z)$ such that for all $x, y \in M$,

$$f(y) - \rho(y-z) \leq F(z) \leq \rho(x+z) - f(x). \tag{21.5}$$

Is there any such number between $f(y)-\rho(y-z)$ and $\rho(x+z)-f(x)$ for every pair $x, y \in M$? This is where $f(x) \leq \rho(x)$ on M and that f is linear is used. For $x, y \in M$,

$$\begin{aligned} &\rho(x+z) - f(x) - [f(y) - \rho(y-z)] \\ &= \rho(x+z) + \rho(y-z) - (f(x) + f(y)) \\ &\geq \rho(x+y) - f(x+y) \geq 0. \end{aligned}$$

Then if $a = \sup\{f(y) - \rho(y-z) : y \in M\}$, $b = \inf\{\rho(x+z) - f(x) : x \in M\}$, it follows that $[a,b] \neq \emptyset$. Choose $F(z)$ in $[a,b]$ so it will satisfy 21.5. This has proved the following lemma.

Lemma 21.2.3 *Let M be a subspace of X, a real linear space, and let ρ be a gauge function on X. Suppose $f : M \to \mathbb{R}$ is linear, $z \notin M$, and $f(x) \leq \rho(x)$ for all $x \in M$. Then f can be extended to $M \oplus \mathbb{R}z$ such that, if F is the extended function, F is linear and $F(x) \leq \rho(x)$ for all $x \in M \oplus \mathbb{R}z$.*

With this lemma, the Hahn Banach theorem can be proved.

Theorem 21.2.4 *(Hahn Banach theorem) Let X be a real vector space, let M be a subspace of X, let $f : M \to \mathbb{R}$ be linear, let ρ be a gauge function on X, and suppose $f(x) \leq \rho(x)$ for all $x \in M$. Then there exists a linear function, $F : X \to \mathbb{R}$, such that*

a.) $F(x) = f(x)$ for all $x \in M$

b.) $F(x) \leq \rho(x)$ for all $x \in X$.

Proof: Let $\mathscr{F} = \{(V,g) : V \supseteq M, V \text{ is a subspace of } X, g : V \to \mathbb{R} \text{ is linear}, g(x) = f(x) \text{ for all } x \in M, \text{ and } g(x) \leq \rho(x) \text{ for } x \in V\}$. Then $(M,f) \in \mathscr{F}$ so $\mathscr{F} \neq \emptyset$. Define a partial order by the following rule. $(V,g) \leq (W,h)$ means

$$V \subseteq W \text{ and } h(x) = g(x) \text{ if } x \in V.$$

By Theorem 21.2.1, there exists a maximal chain, $\mathscr{C} \subseteq \mathscr{F}$. Let $Y = \cup\{V : (V,g) \in \mathscr{C}\}$ and let $h : Y \to \mathbb{R}$ be defined by $h(x) = g(x)$ where $x \in V$ and $(V,g) \in \mathscr{C}$. This is well defined because if $x \in V_1$ and V_2 where (V_1, g_1) and (V_2, g_2) are both in the chain, then since $\mathscr{C}$ is a chain, the two elements are related. Therefore, $g_1(x) = g_2(x)$. Also h is linear because if $ax + by \in Y$, then $x \in V_1$ and $y \in V_2$ where (V_1, g_1) and (V_2, g_2) are elements of $\mathscr{C}$. Therefore, letting V denote the larger of the two V_i, and g be the function that goes with V, it follows $ax + by \in V$ where $(V,g) \in \mathscr{C}$. Therefore,

$$h(ax+by) = g(ax+by) = ag(x) + bg(y) = ah(x) + bh(y).$$

Also, $h(x) = g(x) \leq \rho(x)$ for any $x \in Y$ because for such x, $x \in V$ where $(V,g) \in \mathscr{C}$.

Is $Y = X$? If not, there exists $z \in X \setminus Y$ and there exists an extension of h to $Y \oplus \mathbb{R}z$ using Lemma 21.2.3. Letting $\overline{h}$ denote this extended function, contradicts the maximality of $\mathscr{C}$. Indeed, $\mathscr{C} \cup \{(Y \oplus \mathbb{R}z, \overline{h})\}$ would be a longer chain. ■

This is the original version of the theorem. There is also a version of this theorem for complex vector spaces which is based on a trick.

21.2.3 The Complex Version of the Hahn Banach Theorem

First is a lemma which is quite interesting for its own sake.

Lemma 21.2.5 *Let $h : V \to \mathbb{R}$ where V is a complex normed linear space. Then h is linear with respect to real scalars if and only if $F(x) \equiv h(x) - ih(ix)$ is linear with respect to complex scalars.*

Proof: $\Leftarrow$ By assumption, F is linear with respect to real scalars. Let $c \in \mathbb{R}$. Then

$$cF(x) = ch(x) - cih(ix) = F(cx) = h(cx) - ih(cix).$$

Equating real parts, it follows $ch(x) = h(cx)$. Also

$$\begin{aligned} F(x+y) &= h(x+y) - ih(i(x+y)) = F(x) + F(y) \\ &= h(x) - ih(ix) + h(y) - ih(iy) \end{aligned}$$

Equating real parts, $h(x+y) = h(x) + h(y)$.

$\Rightarrow$ I need to show that F is linear with respect to complex scalars.

$$\begin{aligned} F(ix) &\equiv h(ix) - ih(-x) = h(ix) + ih(x) \\ &= i(h(x) - ih(ix)) = iF(ix) \end{aligned}$$

It is fairly obvious that $F(x+y) = F(x) + F(y)$. Also, if c is real, it is clear that $F(cx) = cF(x)$. Therefore,

$$\begin{aligned} F((a+ib)x) &= F(ax) + F(ibx) \\ &= aF(x) + ibF(x) = (a+ib)F(x) \ \blacksquare \end{aligned}$$

Corollary 21.2.6 *(Hahn Banach) Let M be a subspace of a complex normed linear space X, and suppose $f : M \to \mathbb{C}$ is linear and satisfies $|f(x)| \leq K\|x\|$ for all $x \in M$. Then there exists a linear function F, defined on all of X such that $F(x) = f(x)$ for all $x \in M$ and $|F(x)| \leq K\|x\|$ for all $x \in X$.*

Proof: First note $f(x) = \mathrm{Re}\, f(x) + i\,\mathrm{Im}\, f(x)$ and so

$$\mathrm{Re}\, f(ix) + i\,\mathrm{Im}\, f(ix) = f(ix) = if(x) = i\,\mathrm{Re}\, f(x) - \mathrm{Im}\, f(x).$$

Therefore, $\mathrm{Im}\, f(x) = -\mathrm{Re}\, f(ix)$, and

$$f(x) = \mathrm{Re}\, f(x) - i\,\mathrm{Re}\, f(ix).$$

This is important because it shows it is only necessary to consider $\mathrm{Re}\, f$ in understanding f. From Lemma 21.2.5 $\mathrm{Re}\, f$ is linear with respect to real scalars.

Consider X as a real vector space and let $\rho(x) \equiv K\|x\|$. Then for all $x \in M$,

$$|\mathrm{Re}\, f(x)| \leq |f(x)| \leq K\|x\| \equiv \rho(x).$$

From Theorem 21.2.4, $\mathrm{Re}\, f$ may be extended to a function h which satisfies

$$\begin{aligned} h(ax+by) &= ah(x) + bh(y) \text{ if } a, b \in \mathbb{R} \\ h(x) &\leq K\|x\| \text{ for all } x \in X. \end{aligned}$$

Actually, $|h(x)| \leq K\|x\|$. The reason for this is that $h(-x) = -h(x) \leq K\|-x\| = K\|x\|$ and therefore, $h(x) \geq -K\|x\|$ so $-h(x) \leq K\|x\|$. Thus $|h(x)| \leq K\|x\|$. Let $F(x) \equiv h(x) - ih(ix)$. By Lemma 21.2.5, F is complex linear.

Now $wF(x) = |F(x)|$ for some $|w| = 1$. Therefore

$$\begin{aligned} |F(x)| &= wF(x) = F(wx) \equiv h(wx) - \overbrace{ih(iwx)}^{\text{must equal zero}} = h(wx) \\ &= |h(wx)| \leq K\|wx\| = K\|x\|. \blacksquare \end{aligned}$$

21.2.4 The Dual Space and Adjoint Operators

Definition 21.2.7 *Let X be a Banach space. Denote by X' the space of continuous linear functions which map X to the field of scalars. Thus $X' = \mathscr{L}(X, \mathbb{F})$. By Theorem 21.1.8 on Page 559, X' is a Banach space. Remember with the norm defined on $\mathscr{L}(X, \mathbb{F})$,*

$$\|f\| = \sup\{|f(x)| : \|x\| \leq 1\}$$

X' is called the dual space.

Definition 21.2.8 *Let X and Y be Banach spaces and suppose $L \in \mathscr{L}(X, Y)$. Then define the adjoint map in $\mathscr{L}(Y', X')$, denoted by L^*, by*

$$L^* y^*(x) \equiv y^*(Lx)$$

for all $y^* \in Y'$.

The following diagram is a good one to help remember this definition.

$$\begin{array}{ccc} & L^* & \\ X' & \leftarrow & Y' \\ X & \underset{L}{\rightarrow} & Y \end{array}$$

This is a generalization of the adjoint of a linear transformation on an inner product space from Linear Algebra. Recall

$$(Ax, y) = (x, A^* y)$$

What is being done here is to generalize this algebraic concept to arbitrary Banach spaces. There are some issues which need to be discussed relative to the above definition. First of all, it must be shown that $L^* y^* \in X'$. Also, it will be useful to have the following lemma which is a useful application of the Hahn Banach theorem.

Lemma 21.2.9 *Let X be a normed linear space and let $x \in X \setminus V$ where V is a closed subspace of X. Then there exists $x^* \in X'$ such that $x^*(x) = \|x\| \neq 0$, $x^*(V) = \{0\}$, and $\|x^*\| = \frac{1}{\operatorname{dist}(x,V)}\|x\|$. In the case that $V = \{0\}$, $\|x^*\| = 1$.*

Proof: Let $f : \mathbb{F}x + V \to \mathbb{F}$ be defined by $f(\alpha x + v) = \alpha\|x\|$. First it is necessary to show f is well defined and continuous. If $\alpha_1 x + v_1 = \alpha_2 x + v_2$ then if $\alpha_1 \neq \alpha_2$, then $x \in V$ which is assumed not to happen so f is well defined. It remains to show f is continuous. Suppose

then that $\alpha_n x + v_n \to 0$. It is necessary to show $\alpha_n \to 0$. If this does not happen, then there exists a subsequence, still denoted by α_n such that $|\alpha_n| \geq \delta > 0$. Then $x + (1/\alpha_n) v_n \to 0$. Thus $\|x - (-(1/\alpha_n) v_n)\| \to 0$ so x is a limit of points of V contradicting the assumption that $x \notin V$ and V is a closed subspace. Hence f is continuous on $\mathbb{F}x + V$. Thus

$$\|f\| = \sup_{\|\alpha x+v\|\leq 1} |\alpha| \|x\|.$$

What is $\sup |\alpha|$, given that $\|\alpha x + v\| = |\alpha| \left\|x + \frac{v}{\alpha}\right\| \leq 1$? Since $\frac{v}{\alpha}$ is a generic element of V this reduces to $\sup |\alpha|$ such that $|\alpha| \|x + v\| \leq 1$. Thus the largest $|\alpha|$ can be is $\frac{1}{\operatorname{dist}(x,V)}$ and so

$$\|f\| = \frac{1}{\operatorname{dist}(x,V)} \|x\|$$

Now for $z \in \mathbb{F}x + V, |f(z)| \leq \|f\| \|z\| = \frac{1}{\operatorname{dist}(x,V)} \|x\| \|z\|$. By the complex Hahn Banach theorem, there exists $x^* \in X'$ such that $x^* = f$ on $\mathbb{F}x + V$ and for all $z \in X$,

$$|x^*(z)| \leq \|f\| \|z\| = \frac{1}{\operatorname{dist}(x,V)} \|x\| \|z\|$$

Thus $\|x^*\| \leq \|f\|$. However, equality must occur because

$$\begin{aligned} \frac{1}{\operatorname{dist}(x,V)} \|x\| &= \|f\| \equiv \sup_{\|z\|\leq 1, z\in\mathbb{F}x+V} |f(z)| \\ &= \sup_{\|z\|\leq 1, z\in\mathbb{F}x+V} |x^*(z)| \leq \sup_{\|z\|\leq 1} |x^*(z)| \equiv \|x^*\|. \end{aligned}$$

In case $V = \{0\}, \operatorname{dist}(x,V) = \|x\|$ and so $\|x^*\| = 1$. ■

Theorem 21.2.10 *Let $L \in \mathscr{L}(X,Y)$ where X and Y are Banach spaces. Then*

a.) $L^* \in \mathscr{L}(Y',X')$ as claimed and $\|L^*\| = \|L\|$.

b.) If L maps one to one onto a closed subspace of Y, then L^* is onto.

c.) If L maps onto a dense subset of Y, then L^* is one to one.

Proof: It is routine to verify $L^* y^*$ and L^* are both linear. This follows immediately from the definition. As usual, the interesting thing concerns continuity.

$$\|L^* y^*\| = \sup_{\|x\|\leq 1} |L^* y^*(x)| = \sup_{\|x\|\leq 1} |y^*(Lx)| \leq \|y^*\| \|L\|.$$

Thus L^* is continuous as claimed and $\|L^*\| \leq \|L\|$.

By Lemma 21.2.9, there exists $y_x^* \in Y'$ such that $\|y_x^*\| = 1$ and $y_x^*(Lx) = \|Lx\|$ .Therefore,

$$\begin{aligned} \|L^*\| &= \sup_{\|y^*\|\leq 1} \|L^* y^*\| = \sup_{\|y^*\|\leq 1} \sup_{\|x\|\leq 1} |L^* y^*(x)| \\ &= \sup_{\|y^*\|\leq 1} \sup_{\|x\|\leq 1} |y^*(Lx)| = \sup_{\|x\|\leq 1} \sup_{\|y^*\|\leq 1} |y^*(Lx)| \\ &\geq \sup_{\|x\|\leq 1} |y_x^*(Lx)| = \sup_{\|x\|\leq 1} \|Lx\| = \|L\| \end{aligned}$$

showing that $\|L^*\| \geq \|L\|$ and this shows part a.).

Next consider b.). Let $x^* \in X'$. Is there $y^* \in Y'$ such that $L^*(y^*) = x^*$? This will be so if and only if for all $x \in X, y^*(Lx) = x^*(x)$. Let $f(Lx) \equiv x^*(x)$. This defines f on $L(X)$ because of the assumption that L is one to one. Is f continuous on $L(X)$? Suppose $Lx_n \to Lx$ in $L(X)$. Does it follow that $x_n \to x$? Yes, because $L(X)$ is a closed subspace of a Banach space Y and is therefore also a Banach space. Since L is one to one, it follows from the open mapping theorem that L^{-1} is continuous on $L(X)$ and so indeed, $x_n \to x$ and so $x^*(x_n) \to x^*(x)$ showing that f is continuous on $L(X)$. Now by the Hahn Banach theorem, there is an extension $y^*$of f to all of Y which has the same norm. Thus L^* is onto as claimed.

Consider the last assertion. Suppose $L^*y^* = 0$. Is $y^* = 0$? Letting $Lx \in D$ where D is the dense subset of $Y, y^*(Lx) = L^*y^*(x) = 0$ and so y^* sends all in a dense subset of Y to 0. Hence, by continuity of y^*, it equals 0. Thus L^* is one to one. ■

Corollary 21.2.11 *Suppose X and Y are Banach spaces, $L \in \mathscr{L}(X,Y)$, and L is one to one and onto. Then L^* is also one to one and onto.*

There exists a natural mapping, called the James map from a normed linear space X, to the dual of the dual space which is described in the following definition.

Definition 21.2.12 *Define $J : X \to X''$ by $J(x)(x^*) = x^*(x)$.*

Theorem 21.2.13 *The map J has the following properties.*

a.) J is one to one and linear.

b.) $\|Jx\| = \|x\|$ and $\|J\| = 1$.

c.) $J(X)$ is a closed subspace of X'' if X is complete.

Also if $x^* \in X'$,

$$\|x^*\| = \sup\left\{|x^{**}(x^*)| : \|x^{**}\| \leq 1,\ x^{**} \in X''\right\}.$$

Proof:

$$\begin{aligned} J(ax+by)(x^*) &\equiv x^*(ax+by) = ax^*(x) + bx^*(y) \\ &= (aJ(x) + bJ(y))(x^*). \end{aligned}$$

Since this holds for all $x^* \in X'$, it follows that $J(ax+by) = aJ(x) + bJ(y)$ and so J is linear. If $Jx = 0$, then by Lemma 21.2.9 there exists x^* such that $x^*(x) = \|x\|$ and $\|x^*\| = 1$. Then $0 = J(x)(x^*) = x^*(x) = \|x\|$. This shows a.).

To show b.), let $x \in X$ and use Lemma 21.2.9 to obtain $x^* \in X'$ such that $x^*(x) = \|x\|$ with $\|x^*\| = 1$. Then

$$\begin{aligned} \|x\| &\geq \sup\{|y^*(x)| : \|y^*\| \leq 1\} = \sup\{|J(x)(y^*)| : \|y^*\| \leq 1\} = \|Jx\| \\ &\geq |J(x)(x^*)| = |x^*(x)| = \|x\| \end{aligned}$$

Therefore, $\|Jx\| = \|x\|$ as claimed. Therefore,

$$\|J\| = \sup\{\|Jx\| : \|x\| \leq 1\} = \sup\{\|x\| : \|x\| \leq 1\} = 1.$$

This shows b.).

To verify c.), use b.). If $Jx_n \to y^{**} \in X''$ then by b.), x_n is a Cauchy sequence converging to some $x \in X$ because $\|x_n - x_m\| = \|Jx_n - Jx_m\|$ and $\{Jx_n\}$ is a Cauchy sequence. Then $Jx = \lim_{n\to\infty} Jx_n = y^{**}$.

Finally, to show the assertion about the norm of x^*, use what was just shown applied to the James map from X' to X''' still referred to as J.

$$\|x^*\| = \sup\{|x^*(x)| : \|x\| \leq 1\} = \sup\{|J(x)(x^*)| : \|Jx\| \leq 1\}$$

$$\leq \sup\{|x^{**}(x^*)| : \|x^{**}\| \leq 1\} = \sup\{|J(x^*)(x^{**})| : \|x^{**}\| \leq 1\}$$

$$\equiv \|Jx^*\| = \|x^*\|. \blacksquare$$

Definition 21.2.14 *When J maps X onto X'', X is called reflexive.*

It happens the L^p spaces are reflexive whenever $p > 1$. This is shown later.

21.3 Uniform Convexity of L^p

These terms refer roughly to how round the unit ball is. Here is the definition.

Definition 21.3.1 *A Banach space is uniformly convex if whenever*

$$\|x_n\|,\ \|y_n\| \leq 1$$

and $\|x_n + y_n\| \to 2$, it follows that $\|x_n - y_n\| \to 0$. More precisely, for every $\varepsilon > 0$, there is a $\delta > 0$ such that if $\|x+y\| > 2-\delta$ for $\|x\|, \|y\|$ both no more than 1, then $\|x-y\| < \varepsilon$.

You can show that uniform convexity implies strict convexity. There are various other things which can also be shown. See the exercises for some of these. In this section, it will be shown that the L^p spaces are examples of uniformly convex spaces. This involves some inequalities known as Clarkson's inequalities. Before presenting these, here are the backwards Holder inequality and the backwards Minkowski inequality. Recall that in the Holder inequality, $\frac{p}{p-1} = q = p'$ and for $p > 1$,

$$\int_\Omega |f|\,|g|\,d\mu \leq \left(\int_\Omega |f|^p\,d\mu\right)^{1/p} \left(\int_\Omega |g|^{p/(p-1)}\,d\mu\right)^{(p-1)/p}$$

The idea in these inequalities is to consider the case that $p \in (0,1)$. This inequality is easy to remember if you just take Holder's inequality and turn it around in the case that $0 < p < 1$.

Lemma 21.3.2 *Let $0 < p < 1$ and let f, g be measurable functions . Also*

$$\int_\Omega |g|^{p/(p-1)}\,d\mu < \infty,\ \int_\Omega |f|^p\,d\mu < \infty,$$

which implies that g is 0 only on a set of measure zero. Then the following backwards Holder inequality holds.

$$\int_\Omega |fg|\,d\mu \geq \left(\int_\Omega |f|^p\,d\mu\right)^{1/p} \left(\int_\Omega |g|^{p/(p-1)}\,d\mu\right)^{(p-1)/p}$$

Proof: If $\int |fg|\,d\mu = \infty$, there is nothing to prove. Hence assume this is finite. Then

$$\int |f|^p\,d\mu = \int |g|^{-p}\,|fg|^p\,d\mu$$

This makes sense because, due to the hypothesis on g it must be the case that g equals 0 only on a set of measure zero, since $p/(p-1) < 0$.

Then by the usual Holder inequality, one of the exponents being $1/p > 1$, the other being $1/(1-p)$ also larger than 1 with $p + (1-p) = 1$,

$$\begin{aligned}\int |f|^p\,d\mu &\leq \left(\int |fg|\,d\mu\right)^p \left(\int \left(\frac{1}{|g|^p}\right)^{1/(1-p)} d\mu\right)^{1-p} \\ &= \left(\int |fg|\,d\mu\right)^p \left(\int |g|^{p/p-1}\,d\mu\right)^{1-p}\end{aligned}$$

Now divide by $\left(\int |g|^{p/p-1}\,d\mu\right)^{1-p}$ and then take the p^{th} root. ■

Here is the backwards Minkowski inequality. It looks just like the ordinary Minkowski inequality except the inequality is turned around.

Corollary 21.3.3 *Let $0 < p < 1$ and suppose $\int |h|^p\,d\mu < \infty$ for $h = f, g$. Then*

$$\left(\int (|f| + |g|)^p\,d\mu\right)^{1/p} \geq \left(\int |f|^p\,d\mu\right)^{1/p} + \left(\int |g|^p\,d\mu\right)^{1/p}$$

Proof: If $\int (|f| + |g|)^p\,d\mu = 0$ then there is nothing to prove since this implies $|f| = |g| = 0$ a.e. so assume this is not zero.

$$\int (|f| + |g|)^p\,d\mu = \int (|f| + |g|)^{p-1}\,(|f| + |g|)\,d\mu$$

Since $p < 1$, $(|f| + |g|)^p \leq |f|^p + |g|^p$ and so

$$\int \left((|f| + |g|)^{p-1}\right)^{p/p-1} d\mu < \infty.$$

Hence the backward Holder inequality applies and it follows that

$$\int (|f| + |g|)^p\,d\mu = \int (|f| + |g|)^{p-1}\,|f|\,d\mu + \int (|f| + |g|)^{p-1}\,|g|\,d\mu$$

$$\begin{aligned}&\geq \left(\int \left((|f| + |g|)^{p-1}\right)^{p/p-1} d\mu\right)^{(p-1)/p} \left[\left(\int |f|^p\,d\mu\right)^{1/p} + \left(\int |g|^p\,d\mu\right)^{1/p}\right] \\ &= \left(\int (|f| + |g|)^p\,d\mu\right)^{(p-1)/p} \left[\left(\int |f|^p\,d\mu\right)^{1/p} + \left(\int |g|^p\,d\mu\right)^{1/p}\right]\end{aligned}$$

and so, dividing gives the desired inequality. ■

Consider the "easy" Clarkson inequalities.

Lemma 21.3.4 *For any $p \geq 2$ the following inequality holds for any $t \in [0,1]$,*

$$\left(\frac{1+t}{2}\right)^p + \left(\frac{1-t}{2}\right)^p \leq \frac{1}{2}\left(t^p+1\right)$$

Proof: It is clear that, since $p \geq 2$, the inequality holds for $t = 0$ and $t = 1$.Thus it suffices to consider only $t \in (0,1)$. Let $x = 1/t$. Then, dividing by t^p, the inequality holds if and only if

$$\left(\frac{x+1}{2}\right)^p + \left(\frac{x-1}{2}\right)^p \leq \frac{1}{2}\left(1+x^p\right)$$

for all $x \geq 1$. Let

$$f(x) = \frac{1}{2}\left(1+x^p\right) - \left(\left(\frac{x+1}{2}\right)^p + \left(\frac{x-1}{2}\right)^p\right)$$

Then $f(1) = 0$ and

$$f'(x) = \frac{p}{2}x^{p-1} - \left(\frac{p}{2}\left(\frac{x+1}{2}\right)^{p-1} + \frac{p}{2}\left(\frac{x-1}{2}\right)^{p-1}\right)$$

Since $p-1 \geq 1$, $g(x) = x^{p-1}$ is convex. Its graph is like a smile. Thus $\frac{1}{2}\left(g(x_1)+g(x_2)\right) \geq g\left(\frac{x_1+x_2}{2}\right)$ and so

$$f'(x) \geq \frac{p}{2}x^{p-1} - p\left(\frac{\frac{x+1}{2}+\frac{x-1}{2}}{2}\right)^{p-1} = \frac{p}{2}x^{p-1} - p\left(\frac{x}{2}\right)^{p-1} \geq 0$$

Hence $f(x) \geq 0$ for all $x \geq 1$.■

Corollary 21.3.5 *If $z, w \in \mathbb{C}$ and $p \geq 2$, then*

$$\left|\frac{z+w}{2}\right|^p + \left|\frac{z-w}{2}\right|^p \leq \frac{1}{2}\left(|z|^p + |w|^p\right) \tag{21.6}$$

Proof: One of $|w|, |z|$ is larger. Say $|z| \geq |w|$. Then dividing both sides of the proposed inequality by $|z|^p$ it suffices to verify that for all complex t having $|t| \leq 1$,

$$\left|\frac{1+t}{2}\right|^p + \left|\frac{1-t}{2}\right|^p \leq \frac{1}{2}\left(|t|^p+1\right) \tag{21.7}$$

Say $t = re^{i\theta}$ where $r \leq 1$.Then we need to estimate

$$\left|\frac{1+re^{i\theta}}{2}\right|^p + \left|\frac{1-re^{i\theta}}{2}\right|^p$$

It suffices to show that this is no larger than $\frac{1}{2}\left(r^p+1\right)$. The function on the left in 21.7 equals

$$\begin{aligned} & \frac{1}{2^p}\left(\left(1+r\cos\theta\right)^2 + r^2\sin^2(\theta)\right)^{p/2} + \left(\left(1-r\cos\theta\right)^2 + r^2\sin^2(\theta)\right)^{p/2} \\ = \quad & \frac{1}{2^p}\left(1+r^2+2r\cos\theta\right)^{p/2} + \left(1+r^2-2r\cos\theta\right)^{p/2}, \end{aligned} \tag{21.8}$$

I want to find the maximum value of this function of θ for $\theta \in [0,2\pi]$. By calculus, this will be when the derivative is 0 or at an endpoint. The derivative with respect to θ is $\frac{1}{2^p}$ times

$$\begin{aligned} & \frac{p}{2}\left(\left(1+r^2+2r\cos\theta\right)^{\frac{p-2}{2}}\left(-2r\sin\theta\right)+\left(2r\sin\theta\right)\left(1+r^2-2r\cos\theta\right)^{\frac{p-2}{2}}\right) \\ = & \frac{p}{2}\left(2r\sin\theta\right)\left(\left(1+r^2-2r\cos\theta\right)^{\frac{p-2}{2}}-\left(1+r^2+2r\cos\theta\right)^{\frac{p-2}{2}}\right) \end{aligned}$$

This equals 0 when $\theta = 0, \pi, 2\pi$ or when $\theta = \frac{\pi}{2}, \frac{3\pi}{2}$. At the last two values, the value of the function in 21.8 is

$$\frac{1}{2^{p-1}}\left(1+r^2\right)^{p/2} \leq \frac{1}{2}\left(r^p+1\right).$$

This follows from convexity of $y = x^{p/2}$ for $p \geq 2$. Here is why:

$$\frac{1}{2^{p-1}}\left(1+r^2\right)^{p/2} = \frac{2^{p/2}}{2^{p-1}}\left(\frac{1+r^2}{2}\right)^{p/2} \leq \frac{2^{p/2}}{2^{p-1}}\frac{1}{2}\left(1+r^p\right)$$

At 0 or π, the value of the function in 21.8 is

$$\left(\left(1+r^2-2r\right)^{p/2}+\left(1+r^2+2r\right)^{p/2}\right)\frac{1}{2^p} = \left(\frac{1+r}{2}\right)^p+\left(\frac{1-r}{2}\right)^p$$

and from the above lemma, this is no larger than $\frac{1}{2}\left(r^p+1\right)$. ■

With this corollary, here is the easy Clarkson inequality.

Theorem 21.3.6 *Let $p \geq 2$. Then*

$$\left\|\frac{f+g}{2}\right\|_{L^p}^p+\left\|\frac{f-g}{2}\right\|_{L^p}^p \leq \frac{1}{2}\left(\|f\|_{L^p}^p+\|g\|_{L^p}^p\right)$$

Proof: This follows right away from the above corollary.

$$\int_\Omega\left|\frac{f+g}{2}\right|^p d\mu+\int_\Omega\left|\frac{f-g}{2}\right|^p d\mu \leq \frac{1}{2}\int_\Omega\left(|f|^p+|g|^p\right)d\mu \ ■$$

Now it remains to consider the hard Clarkson inequalities. These pertain to $p < 2$. First is the following elementary inequality.

Lemma 21.3.7 *For $1 < p < 2$, the following inequality holds for all $t \in [0,1]$.*

$$\left(\frac{1+t}{2}\right)^q+\left(\frac{1-t}{2}\right)^q \leq \left(\frac{1}{2}+\frac{1}{2}t^p\right)^{q/p}$$

where here $1/p+1/q = 1$ so $q > 2$.

Proof: First note that if $t = 0$ or 1, the inequality holds. Next observe that the map $s \to \frac{1-s}{1+s}$ maps $(0,1)$ onto $(0,1)$. Replace t with $(1-s)/(1+s)$. Then the desired inequality is equivalent to the following for $s \in (0,1)$.

$$\left(\frac{1}{s+1}\right)^q+\left(\frac{s}{s+1}\right)^q \leq \left(\frac{1}{2}+\frac{1}{2}\left(\frac{1-s}{s+1}\right)^p\right)^{q/p}$$

Multiplying both sides by $(1+s)^q$, this inequality is equivalent to showing that for all $s \in (0,1)$,

$$\begin{aligned} 1+s^q &\leq ((1+s)^p)^{q/p}\left(\frac{1}{2}+\frac{1}{2}\left(\frac{1-s}{s+1}\right)^p\right)^{q/p} \\ &= \left(\frac{1}{2}\right)^{q/p}((1+s)^p+(1-s)^p)^{q/p} \end{aligned}$$

This is the same as establishing

$$\frac{1}{2}((1+s)^p+(1-s)^p)-(1+s^q)^{p-1} \geq 0 \tag{21.9}$$

where $p-1 = p/q$ due to the definition of q above. Note how this has reduced to an expression in which exponents are p or $p-1$ rather than q. We know $p \in (1,2)$ whereas, q is something larger than 2.

$$\left(\begin{array}{c} p \\ l \end{array}\right) \equiv \frac{p(p-1)\cdots(p-l+1)}{l!},\ l \geq 1$$

and $\left(\begin{array}{c} p \\ 0 \end{array}\right) \equiv 1$. What is the sign of $\left(\begin{array}{c} p \\ l \end{array}\right)$? Recall that $1 < p < 2$ so the sign is positive if $l = 0, l = 1, l = 2$. What about $l = 3$? $\left(\begin{array}{c} p \\ 3 \end{array}\right) = \frac{p(p-1)(p-2)}{3!}$ so this is negative. Then $\left(\begin{array}{c} p \\ 4 \end{array}\right)$ is positive. Thus these alternate between positive and negative with $\left(\begin{array}{c} p \\ 2k \end{array}\right) > 0$ for all k. What about $\left(\begin{array}{c} p-1 \\ k \end{array}\right)$? When $k = 0$ it is positive. When $k = 1$ it is also positive. When $k = 2$ it equals $\frac{(p-1)(p-2)}{2!} < 0$. Then when $k = 3$, $\left(\begin{array}{c} p-1 \\ 3 \end{array}\right) > 0$. Thus $\left(\begin{array}{c} p-1 \\ k \end{array}\right)$ is positive when k is odd and is negative when k is even.

Now return to 21.9. The left side equals

$$\frac{1}{2}\left(\sum_{k=0}^{\infty}\left(\begin{array}{c} p \\ k \end{array}\right)s^k+\sum_{k=0}^{\infty}\left(\begin{array}{c} p \\ k \end{array}\right)(-s)^k\right)-\sum_{k=0}^{\infty}\left(\begin{array}{c} p-1 \\ k \end{array}\right)s^{qk}.$$

The first term equals 0. Then this reduces to

$$\sum_{k=1}^{\infty}\left(\begin{array}{c} p \\ 2k \end{array}\right)s^{2k}-\left(\begin{array}{c} p-1 \\ 2k \end{array}\right)s^{q2k}-\left(\begin{array}{c} p-1 \\ 2k-1 \end{array}\right)s^{q(2k-1)}$$

From the above observation about the binomial coefficients, the above is larger than

$$\sum_{k=1}^{\infty}\left(\begin{array}{c} p \\ 2k \end{array}\right)s^{2k}-\left(\begin{array}{c} p-1 \\ 2k-1 \end{array}\right)s^{q(2k-1)}$$

It remains to show the k^{th} term in the above sum is nonnegative. Now $q(2k-1) > 2k$ for all $k \geq 1$ because $q > 2$. Then since $0 < s < 1$

$$\left(\begin{array}{c} p \\ 2k \end{array}\right)s^{2k}-\left(\begin{array}{c} p-1 \\ 2k-1 \end{array}\right)s^{q(2k-1)} \geq s^{2k}\left(\left(\begin{array}{c} p \\ 2k \end{array}\right)-\left(\begin{array}{c} p-1 \\ 2k-1 \end{array}\right)\right)$$

However, this is nonnegative because it equals

$$\begin{aligned} & s^{2k}\left(\frac{p(p-1)\cdots(p-2k+1)}{(2k)!}-\frac{\overbrace{(p-1)(p-2)\cdots(p-2k+1)}^{>0}}{(2k-1)!}\right) \\ \geq\ & s^{2k}\left(\frac{p(p-1)\cdots(p-2k+1)}{(2k)!}-\frac{(p-1)(p-2)\cdots(p-2k+1)}{(2k)!}\right) \\ =\ & s^{2k}\frac{(p-1)(p-2)\cdots(p-2k+1)}{(2k)!}(p-1)>0. \blacksquare \end{aligned}$$

Corollary 21.3.8 *Let $z,w\in\mathbb{C}$. Then for $p\in(1,2)$,*

$$\left|\frac{z+w}{2}\right|^q+\left|\frac{z-w}{2}\right|^q\leq\left(\frac{1}{2}|z|^p+\frac{1}{2}|w|^p\right)^{q/p}$$

Proof: One of $|w|,|z|$ is larger. Say $|w|\geq|z|$. Then dividing by $|w|^q$, for $t=z/w$, showing the above inequality is equivalent to showing that for all $t\in\mathbb{C}$, $|t|\leq 1$,

$$\left|\frac{t+1}{2}\right|^q+\left|\frac{1-t}{2}\right|^q\leq\left(\frac{1}{2}|t|^p+\frac{1}{2}\right)^{q/p}$$

Now $q>2$ and so by the same argument given in proving Corollary 21.3.5, for $t=re^{i\theta}$, the left side of the above inequality is maximized when $\theta=0$. Hence, from Lemma 21.3.7,

$$\begin{aligned} \left|\frac{t+1}{2}\right|^q+\left|\frac{1-t}{2}\right|^q &\leq \left|\frac{|t|+1}{2}\right|^q+\left|\frac{1-|t|}{2}\right|^q \\ &\leq\left(\frac{1}{2}|t|^p+\frac{1}{2}\right)^{q/p}. \blacksquare \end{aligned}$$

From this the hard Clarkson inequality follows. The two Clarkson inequalities are summarized in the following theorem.

Theorem 21.3.9

Let $2\leq p$. Then

$$\left\|\frac{f+g}{2}\right\|_{L^p}^p+\left\|\frac{f-g}{2}\right\|_{L^p}^p\leq\frac{1}{2}\left(\|f\|_{L^p}^p+\|g\|_{L^p}^p\right)$$

Let $1<p<2$. Then for $1/p+1/q=1$,

$$\left\|\frac{f+g}{2}\right\|_{L^p}^q+\left\|\frac{f-g}{2}\right\|_{L^p}^q\leq\left(\frac{1}{2}\|f\|_{L^p}^p+\frac{1}{2}\|g\|_{L^p}^p\right)^{q/p}$$

Proof: The first was established above. Consider the second.

$$\left\|\frac{f+g}{2}\right\|_{L^p}^q+\left\|\frac{f-g}{2}\right\|_{L^p}^q=$$

$$\left(\int_\Omega\left|\frac{f+g}{2}\right|^p d\mu\right)^{q/p}+\left(\int_\Omega\left|\frac{f-g}{2}\right|^p d\mu\right)^{q/p}$$

$$= \left(\int_\Omega \left(\left| \frac{f+g}{2} \right|^q \right)^{p/q} d\mu \right)^{q/p} + \left(\int_\Omega \left(\left| \frac{f-g}{2} \right|^q \right)^{p/q} d\mu \right)^{q/p}$$

Now $p/q < 1$ and so the backwards Minkowski inequality applies. Thus

$$\leq \left(\int_\Omega \left(\left| \frac{f+g}{2} \right|^q + \left| \frac{f-g}{2} \right|^q \right)^{p/q} d\mu \right)^{q/p}$$

From Corollary 21.3.8,

$$\begin{aligned} &\leq \left(\int_\Omega \left(\left(\frac{1}{2}|f|^p + \frac{1}{2}|g|^p \right)^{q/p} \right)^{p/q} d\mu \right)^{q/p} \\ &= \left(\int_\Omega \left(\frac{1}{2}|f|^p + \frac{1}{2}|g|^p \right) d\mu \right)^{q/p} = \left(\frac{1}{2}\|f\|_{L^p}^p + \frac{1}{2}\|g\|_{L^p}^p \right)^{q/p} \blacksquare \end{aligned}$$

Now with these Clarkson inequalities, it is not hard to show that all the L^p spaces are uniformly convex.

Theorem 21.3.10 *The L^p spaces are uniformly convex.*

Proof: First suppose $p \geq 2$. Suppose $\|f_n\|_{L^p}, \|g_n\|_{L^p} \leq 1$ and $\left\| \frac{f_n+g_n}{2} \right\|_{L^p} \to 1$. Then from the first Clarkson inequality,

$$\left\| \frac{f_n+g_n}{2} \right\|_{L^p}^p + \left\| \frac{f_n-g_n}{2} \right\|_{L^p}^p \leq \frac{1}{2}\left(\|f_n\|_{L^p}^p + \|g_n\|_{L^p}^p \right) \leq 1$$

and so $\|f_n - g_n\|_{L^p} \to 0$.

Next suppose $1 < p < 2$ and $\left\| \frac{f_n+g_n}{2} \right\|_{L^p} \to 1$. Then from the second Clarkson inequality

$$\left\| \frac{f_n+g_n}{2} \right\|_{L^p}^q + \left\| \frac{f_n-g_n}{2} \right\|_{L^p}^q \leq \left(\frac{1}{2}\|f_n\|_{L^p}^p + \frac{1}{2}\|g_n\|_{L^p}^p \right)^{q/p} \leq 1$$

which shows that $\|f_n - g_n\|_{L^p} \to 0$. $\blacksquare$

21.4 Closed Subspaces

Theorem 21.4.1 *Let X be a Banach space and let $V = \operatorname{span}(x_1, \cdots, x_n)$. Then V is a closed subspace of X.*

Proof: Without loss of generality, it can be assumed $\{x_1, \cdots, x_n\}$ is linearly independent. Otherwise, delete those vectors which are in the span of the others till a linearly independent set is obtained. Let

$$x = \lim_{p\to\infty} \sum_{k=1}^n c_k^p x_k \in \overline{V}. \tag{21.10}$$

First suppose $\boldsymbol{c}^p \equiv \left(c_1^p, \cdots, c_n^p\right)$ is not bounded in $\mathbb{F}^n$. Then $\boldsymbol{d}^p \equiv \boldsymbol{c}^p / |\boldsymbol{c}^p|_{\mathbb{F}^n}$ is a unit vector in $\mathbb{F}^n$ and so there exists a subsequence, still denoted by $\boldsymbol{d}^p$ which converges to $\boldsymbol{d}$ where $|\boldsymbol{d}| = 1$. Then

$$\mathbf{0} = \lim_{p\to\infty} \frac{x}{\|\boldsymbol{c}^p\|} = \lim_{p\to\infty} \sum_{k=1}^{n} d_k^p x_k = \sum_{k=1}^{n} d_k x_k$$

where $\sum_k |d_k|^2 = 1$ in contradiction to the linear independence of the $\{x_1, \cdots, x_n\}$. Hence it must be the case that $\boldsymbol{c}^p$ is bounded in $\mathbb{F}^n$. Then taking a subsequence, still denoted as p, it can be assumed $\boldsymbol{c}^p \to \boldsymbol{c}$ and then in 21.10 it follows $x = \sum_{k=1}^{n} c_k x_k \in \operatorname{span}(x_1, \cdots, x_n)$. ■

Proposition 21.4.2 *Let E be a separable Banach space. Then there exists an increasing sequence of subspaces, $\{F_n\}$ such that $\dim(F_{n+1}) - \dim(F_n) \leq 1$ and equals 1 for all n if the dimension of E is infinite. Also $\cup_{n=1}^{\infty} F_n$ is dense in E. In the case where E is infinite dimensional, $F_n = \operatorname{span}(e_1, \cdots, e_n)$ where for each n*

$$\operatorname{dist}(e_{n+1}, F_n) \geq \frac{1}{2} \tag{21.11}$$

and defining,

$$G_k \equiv \operatorname{span}\left(\{e_j : j \neq k\}\right)$$

$$\operatorname{dist}(e_k, G_k) \geq \frac{1}{4}. \tag{21.12}$$

Proof: Since E is separable, so is $\partial B(0,1)$, the boundary of the unit ball thanks to Corollary 3.4.3. Let $\{w_k\}_{k=1}^{\infty}$ be a countable dense subset of $\partial B(0,1)$.

Let $e_1 = w_1$. Let $F_1 = \mathbb{F}e_1$. Suppose F_n has been obtained and equals the following: $\operatorname{span}(e_1, \cdots, e_n)$ where $\{e_1, \cdots, e_n\}$ is independent, $\|e_k\| = 1$, and

$$\operatorname{dist}(e_n, \operatorname{span}(e_1, \cdots, e_{n-1})) \geq \frac{1}{2}.$$

For each n, F_n is closed by Theorem 21.4.1.

If F_n contains $\{w_k\}_{k=1}^{\infty}$, let $F_m = F_n$ for all $m > n$. Otherwise, pick $w \in \{w_k\}$ to be the point of $\{w_k\}_{k=1}^{\infty}$ having the smallest subscript which is not contained in F_n. Then w is at a positive distance λ from F_n because F_n is closed. Therefore, there exists $y \in F_n$ such that $\lambda \leq \|y - w\| \leq 2\lambda$. Let $e_{n+1} = \frac{w-y}{\|w-y\|}$. It follows

$$w = \|w - y\| e_{n+1} + y \in \operatorname{span}(e_1, \cdots, e_{n+1}) \equiv F_{n+1}$$

Then if $x \in \operatorname{span}(e_1, \cdots, e_n)$,

$$\begin{aligned} \|e_{n+1} - x\| &= \left\| \frac{w-y}{\|w-y\|} - x \right\| = \left\| \frac{w-y}{\|w-y\|} - \frac{\|w-y\| x}{\|w-y\|} \right\| \\ &\geq \frac{1}{2\lambda} \|w - y - \|w-y\| x\| \geq \frac{\lambda}{2\lambda} = \frac{1}{2}. \end{aligned}$$

This has shown the existence of an increasing sequence of subspaces, $\{F_n\}$ as described above. It remains to show the union of these subspaces is dense. First note that the union of these subspaces must contain the $\{w_k\}_{k=1}^{\infty}$ because if w_m is missing, then it would contradict

the construction at the m^{th} step. That one should have been chosen. However, $\{w_k\}_{k=1}^{\infty}$ is dense in $\partial B(0,1)$. If $x \in E$ and $x \neq 0$, then $\frac{x}{\|x\|} \in \partial B(0,1)$ then there exists

$$w_m \in \{w_k\}_{k=1}^{\infty} \subseteq \cup_{n=1}^{\infty} F_n$$

such that $\left\| w_m - \frac{x}{\|x\|} \right\| < \frac{\varepsilon}{\|x\|}$. But then $\| \|x\| w_m - x \| < \varepsilon$. and so $\|x\| w_m$ is a point of $\cup_{n=1}^{\infty} F_n$ which is within ε of x. This proves $\cup_{n=1}^{\infty} F_n$ is dense as desired. 21.11 follows from the construction. It remains to verify 21.12.

Let $y \in G_k$. Thus for some $n, y = \sum_{j=1}^{k-1} c_j e_j + \sum_{j=k+1}^{n} c_j e_j$ and I need to show $\|y - e_k\| \geq 1/4$. Without loss of generality, $c_n \neq 0$ and $n > k$. Suppose 21.12 does not hold for some such y so that

$$\left\| e_k - \left(\sum_{j=1}^{k-1} c_j e_j + \sum_{j=k+1}^{n} c_j e_j \right) \right\| < \frac{1}{4}. \tag{21.13}$$

Then from the construction,

$$\frac{1}{4} > |c_n| \left\| e_k - \left(\sum_{j=1}^{k-1} (c_j/c_n) e_j + \sum_{j=k+1}^{n-1} (c_j/c_n) e_j + e_n \right) \right\| \geq |c_n| \frac{1}{2}$$

and so $|c_n| < 1/2$. Consider the left side of 21.13. By the construction

$$\begin{aligned} \frac{1}{4} &> \left\| \overbrace{c_n (e_k - e_n) + (1 - c_n) e_k}^{e_k - c_n e_n} - \left(\sum_{j=1}^{k-1} c_j e_j + \sum_{j=k+1}^{n-1} c_j e_j \right) \right\| \\ &\geq |1 - c_n| - |c_n| \left\| (e_k - e_n) - \left(\sum_{j=1}^{k-1} (c_j/c_n) e_j + \sum_{j=k+1}^{n-1} (c_j/c_n) e_j \right) \right\| \\ &\geq |1 - c_n| - |c_n| \frac{1}{2} \geq 1 - \frac{3}{2} |c_n| > 1 - \frac{3}{2} \frac{1}{2} = \frac{1}{4}, \end{aligned}$$

a contradiction. This proves the desired estimate. ■

Definition 21.4.3 *A Banach space X has a Schauder basis $\{e_k\}_{k=1}^{\infty}$ if for every $x \in X$, there are unique scalars c_k such that $x = \sum_{k=1}^{\infty} c_k x_k$. This is different than a basis because you allow countable sums. For example, you might consider Fourier series.*

21.5 Weak And Weak $*$ Topologies

Proposition 21.4.2 shows that in infinite dimensional space, closed and bounded will not be compact. However, in applications one would like to be able to get convergence of subsequences. This involves asking for less than norm convergence and the concept of weak topologies.

21.5.1 Basic Definitions

Let X be a Banach space and let X' be its dual space.[1] For A' a **finite** subset of X', denote by $\rho_{A'}$ the function defined on X

$$\rho_{A'}(x) \equiv \max_{x^* \in A'} |x^*(x)| \tag{21.14}$$

[1] Actually, all this works in much more general settings than this.

and also let $B_{A'}(x,r)$ be defined by

$$B_{A'}(x,r) \equiv \{y \in X : \rho_{A'}(y-x) < r\} \tag{21.15}$$

Then certain things are obvious. First of all, if $a \in \mathbb{F}$ and $x, y \in X$,

$$\begin{aligned} \rho_{A'}(x+y) &\leq \rho_{A'}(x) + \rho_{A'}(y), \\ \rho_{A'}(ax) &= |a| \rho_{A'}(x). \end{aligned}$$

Similarly, letting A be a finite subset of X, denote by ρ_A the function defined on X'

$$\rho_A(x^*) \equiv \max_{x \in A} |x^*(x)| \tag{21.16}$$

and let $B_A(x^*, r)$ be defined by

$$B_A(x^*, r) \equiv \left\{y^* \in X' : \rho_A(y^* - x^*) < r\right\}. \tag{21.17}$$

It is also clear that

$$\begin{aligned} \rho_A(x^* + y^*) &\leq \rho(x^*) + \rho_A(y^*), \\ \rho_A(ax^*) &= |a| \rho_A(x^*). \end{aligned}$$

Lemma 21.5.1 *The sets $B_{A'}(x,r)$ where A' is a finite subset of X' and $x \in X$ form a basis for a topology on X known as the weak topology. The sets $B_A(x^*, r)$ where A is a finite subset of X and $x^* \in X'$ form a basis for a topology on X' known as the weak $*$ topology.*

Proof: The two assertions are very similar. I will verify the one for the weak topology. The union of these sets, $B_{A'}(x,r)$ for $x \in X$ and $r > 0$ is all of X. Now suppose z is contained in the intersection of two of these sets. Say

$$z \in B_{A'}(x,r) \cap B_{A'_1}(x_1, r_1)$$

Then let $C' = A' \cup A'_1$ and let

$$0 < \delta \leq \min\left(r - \rho_{A'}(z-x), r_1 - \rho_{A'_1}(z-x_1)\right).$$

Consider $y \in B_{C'}(z, \delta)$. Then

$$r - \rho_{A'}(z-x) \geq \delta > \rho_{C'}(y-z) \geq \rho_{A'}(y-z)$$

and so

$$r > \rho_{A'}(y-z) + \rho_{A'}(z-x) \geq \rho_{A'}(y-x)$$

which shows $y \in B_{A'}(x,r)$. Similar reasoning shows $y \in B_{A'_1}(x_1, r_1)$ and so

$$B_{C'}(z, \delta) \subseteq B_{A'}(x,r) \cap B_{A'_1}(x_1, r_1).$$

Therefore, these sets are a basis for a toplogy known as the weak topology which consists of the union of all sets of the form $B_{A'}(x,r)$. ■

21.5.2 Banach Alaoglu Theorem

Why does anyone care about these topologies? The short answer is that in the weak $*$ topology, the closed unit ball in X' is compact. This is not true in the norm topology thanks to Proposition 21.4.2. This wonderful result is the Banach Alaoglu theorem. First recall the notion of the product topology, and the Tychonoff theorem, Theorem 19.3.2 on Page 525 which are stated here for convenience.

Definition 21.5.2 *Let I be a set and suppose for each $i \in I$, (X_i, τ_i) is a nonempty topological space. The Cartesian product of the X_i, denoted by $\prod_{i\in I} X_i$, consists of the set of all choice functions defined on I which select a single element of each X_i. Thus $f \in \prod_{i\in I} X_i$ means for every $i \in I$, $f(i) \in X_i$. The axiom of choice says $\prod_{i\in I} X_i$ is nonempty. Next is a description of a subbasis for a topology. Let $P_j(A) = \prod_{i\in I} B_i$ where $B_i = X_i$ if $i \neq j$ and $B_j = A$. A subbasis for a topology on the product space consists of all sets $P_j(A)$ where $A \in \tau_j$. (These sets have an open set from the topology of X_j in the j^{th} slot and the whole space in the other slots.) Thus a basis consists of finite intersections of these sets. Note that the intersection of two of these basic sets is another basic set and their union yields $\prod_{i\in I} X_i$. Therefore, they satisfy the condition needed for a collection of sets to serve as a basis for a topology. This topology is called the product topology and is denoted by $\prod \tau_i$.*

Theorem 21.5.3 *If (X_i, τ_i) is compact, then so is $(\prod_{i\in I} X_i, \prod \tau_i)$.*

The Banach Alaoglu theorem is as follows.

Theorem 21.5.4 *Let B' be the closed unit ball in X'. Then B' is compact in the weak $*$ topology.*

Proof: By the Tychonoff theorem, Theorem 21.5.3, $P \equiv \prod_{x\in X} \overline{B(0,\|x\|)}$ is compact in the product topology where the topology on $\overline{B(0,\|x\|)}$ is the usual topology of $\mathbb{F}$. Recall P is the set of functions which map a point $x \in X$ to a point in $\overline{B(0,\|x\|)}$. Therefore, $B' \subseteq P$. Also the basic open sets in the weak $*$ topology on B' are obtained as the intersection of basic open sets in the product topology of P to B' and so it suffices to show B' is a closed subset of P. Suppose then that $f \in P \setminus B'$. Since $|f(x)| \leq \|x\|$ for each x, it follows f cannot be linear. There are two ways this can happen. One way is that for some x, y

$$f(x+y) \neq f(x) + f(y) \tag{21.18}$$

for some $x, y \in X$ and the other is that $f(\lambda x) \neq \lambda f(x)$ for some λ, x. Consider the first. If g is close enough to f at the three points, $x+y, x$, and y, 21.18 will hold for g in place of f. In other words there is a basic open set containing f, such that for all g in this basic open set, $g \notin B'$. A similar consideration applies in case $f(\lambda x) \neq \lambda f(x)$ for some scalar λ and x. Since $P \setminus B'$ is open, it follows B' is a closed subset of P and is therefore, compact. ■

Note that if the canonical map $J : X \to X''$ discussed earlier given by $Jx(x^*) \equiv x^*(x)$ is onto, then we could conclude that B the closed unit ball in X is weakly compact because J would be a homeomorphism of B'' and B. Thus reflexive spaces are important in these considerations.

Sometimes one can consider the weak $*$ topology as a metric space. You can do it for K when K is weak $*$ compact and X is separable. Note that it was just shown that the closed ball is weak $*$ compact.

Theorem 21.5.5 *If $K \subseteq X'$ is compact in the weak ∗ topology and X is separable in the weak topology then there exists a metric d, on K such that if τ_d is the topology on K induced by d and if τ is the topology on K induced by the weak ∗ topology of X', then $\tau = \tau_d$. Thus one can consider K with the weak ∗ topology as a metric space.*

Proof: Let $D = \{x_n\}$ be the dense countable subset in X. The metric is

$$d(f,g) \equiv \sum_{n=1}^{\infty} 2^{-n} \frac{\rho_{x_n}(f-g)}{1+\rho_{x_n}(f-g)}$$

where $\rho_{x_n}(f) = |f(x_n)|$. Clearly $d(f,g) = d(g,f) \geq 0$. If $d(f,g) = 0$, then this requires $f(x_n) = g(x_n)$ for all $x_n \in D$. Is it the case that $f = g$? Does $f(x) = g(x)$ for all x, not just for the x_n?

Letting x be given, $B_{\{f,g\}}(x,r)$ contains some $x_n \in D$. Hence

$$\max\{|f(x_n) - f(x)|, |g(x_n) - g(x)|\} < r$$

and $f(x_n) = g(x_n)$. It follows that $|f(x) - g(x)| \leq$

$$|f(x) - f(x_n)| + + \left| \overbrace{f(x_n) - g(x_n)}^{=0} \right| + |g(x_n) - g(x)| < 2r.$$

Since r is arbitrary, this implies $f(x) = g(x)$.

It is routine to verify the triangle inequality from the easy to establish inequality,

$$\frac{x}{1+x} + \frac{y}{1+y} \geq \frac{x+y}{1+x+y},$$

valid whenever $x, y \geq 0$. Therefore this is a metric.

Thus there are two topological spaces, (K,τ) and (K,d), the first being K with the weak ∗ topology and the second being K with the topology from this metric. Suppose $B(f,r)$ is an open ball with respect to the metric space topology. I claim that $B(f,r)$ is open in the weak ∗ topology τ. To do this, let $d(f,g) < r$. Is there a finite set $A \subseteq X$ such that $B_A(g,\delta) \subseteq B(f,r)$? Let $A_n \equiv \{x_1, ..., x_n\}$ and pick n large enough that $\sum_{k=n}^{\infty} 2^{-n} < \frac{r-d(f,g)}{2} \equiv \delta$. Then if $h \in B_{A_n}(g,\delta)$, it follows that

$$\begin{aligned} d(f,h) &\leq d(f,g) + d(g,h) < d(f,g) + \sum_{k=1}^{n-1} 2^{-k} |g(x_k) - h(x_k)| + \frac{r - d(f,g)}{2} \\ &< d(f,g) + \sum_{k=1}^{\infty} \delta 2^{-k} + \delta = d(f,g) + 2\delta = d(f,g) + (r - d(f,g)) = r \end{aligned}$$

Thus $B_A(g,\delta) \subseteq B(f,r)$ and so $B(f,r)$ is the union of weak ∗ open sets and is therefore, weakly open. It follows that $\tau_d \subseteq \tau$. Thus it is clear that if i is the identity map, $i : (K,\tau) \to (K,d)$, then i is continuous.

Now suppose $U \in \tau$. Is U in τ_d? Since K is compact with respect to τ, it follows from the above that K is compact with respect to $\tau_d \subseteq \tau$. Hence $K \setminus U$ is compact with respect to τ_d and so it is closed with respect to τ_d. Thus U is open with respect to τ_d. The identity map $i : (K,d) \to (K,\tau)$ is continuous. ■

Note that the above proof is about the elements of X' continuous with respect to the weak topology on X and D is dense with respect to this **weak topology**.

The fact that this set with the weak $*$ topology can be considered a metric space is very significant because if a point is a limit point in a metric space, one can extract a convergent sequence. Also this has shown that the closed unit ball in X' can be considered a metric space provided X is separable. Therefore, it is sequentially compact from Theorem 3.5.8.

Note that if a Banach space is separable, then it is weakly separable. In fact, the countable dense set D with respect to the norm is also dense with respect to the weak topology. To see this, suppose $A = \{x_1^*, ...x_m^*\}$ is a finite subset of X'. Consider $B_A(x,r)$. Is there a point of D in $B_A(x,r)$? Choose $x_n \in D$ close enough to x that

$$\max\{|x_i^*(x) - x_i^*(x_n)|, i = 1,...,m\} < r$$

This can be done because $|x_i^*(x) - x_i^*(x_n)| \leq \|x_i^*\| \|x - x_n\|$ and this can be made small by taking

$$\|x - x_n\| < \min\left\{\frac{r}{1+\|x_i^*\|}, i = 1,...,m\right\}.$$

Corollary 21.5.6 *If X is weakly separable and $K \subseteq X'$ is compact in the weak $*$ topology, then K is sequentially compact. That is, if $\{f_n\}_{n=1}^{\infty} \subseteq K$, then there exists a subsequence f_{n_k} and $f \in K$ such that for all $x \in X$, $\lim_{k\to\infty} f_{n_k}(x) = f(x)$.*

Proof: By Theorem 21.5.5, K is a metric space for the metric described there and it is compact. Therefore by the characterization of compact metric spaces, Proposition 3.5.8 on Page 78, K is sequentially compact. This proves the corollary. ■

21.5.3 Eberlein Smulian Theorem

Next consider the weak topology. The most interesting results have to do with a reflexive Banach space. The following lemma ties together the weak and weak $*$ topologies in the case of a reflexive Banach space. It shows that a reflexive Banach space is actually a dual space.

Definition 21.5.7 *For X a Banach space, define $J : X \to X''$ by: For $x^* \in X'$, $Jx(x^*) \equiv x^*(x)$.*

For the properties of J see Theorem 21.2.13.

Lemma 21.5.8 *Let $J : X \to X''$ be the James map $Jx(x^*) \equiv x^*(x)$ and let X be reflexive so that J is onto. Then $J : (X,$ weak topology$) \to (X'',$ weak $*$ topology$)$ is a homeomorphism. This means J is one to one, onto, and both J and J^{-1} are continuous.*

Proof: Let $x^* \in X'$ and let $B_{x^*}(x,r) \equiv \{y : |x^*(x) - x^*(y)| < r\}$. Thus $B_{x^*}(x,r)$ is a subbasic set for the weak topology on X. I claim that $JB_{x^*}(x,r) = B_{x^*}(Jx,r)$. where $B_{x^*}(Jx,r)$ is a subbasic set for the weak $*$ topology on X''. If $y \in B_{x^*}(x,r)$, then $\|Jy - Jx\| = \|x - y\| < r$ and so $JB_{x^*}(x,r) \subseteq B_{x^*}(Jx,r)$. Now if $x^{**} \in B_{x^*}(Jx,r)$, then since J is reflexive, there exists $y \in X$ such that $Jy = x^{**}$ and so $\|y - x\| = \|Jy - Jx\| < r$ showing that $JB_{x^*}(x,r) = B_{x^*}(Jx,r)$. A typical subbasic set in the weak $*$ topology is of the form $B_{x^*}(Jx,r)$. Thus J maps the subbasic sets of the weak topology to the subbasic sets of the weak $*$ topology of X''. Therefore, J is a homeomorphism as claimed. ■

The following is an easy corollary.

Corollary 21.5.9 *If X is a reflexive Banach space, then the closed unit ball is weakly compact.*

Proof: Let B be the closed unit ball. Then $B = J^{-1}(B^{**})$ where B^{**} is the unit ball in X'' which is compact in the weak $*$ topology. Therefore B is weakly compact because J^{-1} is continuous. ■

Corollary 21.5.10 *If X is a reflexive Banach space, and X' is weak $*$ separable, then the closed unit ball is weakly sequentially compact.*

Proof: This follows from Corollary 21.5.6 because X can be considered as the dual space of X' according to the rule $x(x^*) \equiv x^*(x)$. This definition gives x as continuous and linear. It is clearly linear and is continuous because $|x(x^*)| = |x^*(x)| \leq \|x^*\| \|x\|$ If $f \in X''$, then $f = Jx$ and so $f(x^*) = Jx(x^*) \equiv x^*(x) \equiv x(x^*)$. Since X' is weakly separable, it follows from the above corollary. To reiterate the reasoning, B is the unit ball in X'' and so it is weak $*$ compact and is also a metric space so it is weakly sequentially compact also. ■

In fact if $K \subseteq X$ is weakly compact and X is reflexive with X' separable, then K is sequentially weakly compact.

Corollary 21.5.11 *Let X be a reflexive Banach space. If $K \subseteq X$ is compact in the weak topology and X' is separable in the weak $*$ topology, then there exists a metric d, on K such that if τ_d is the topology on K induced by d and if τ is the topology on K induced by the weak topology of X, then $\tau = \tau_d$. Thus one can consider K with the weak topology as a metric space. Thus K is weakly sequentially compact.*

Proof: This follows from Theorem 21.5.5 and Lemma 21.5.8. Lemma 21.5.8 implies $J(K)$ is compact in X''. Then since X' is separable in the weak $*$ topology, X is separable in the weak topology and so there is a metric, d'' on $J(K)$ which delivers the weak $*$ topology on $J(K)$. Let $d(x,y) \equiv d''(Jx, Jy)$. Then

$$(K, \tau_d) \xrightarrow{J} (J(K), \tau_{d''}) \xrightarrow{id} (J(K), \tau_{\text{weak}\,*}) \xrightarrow{J^{-1}} (K, \tau_{\text{weak}})$$

and all the maps are homeomorphisms. ■

Recall Lemma 21.2.9.

Lemma 21.5.12 *Let Y be a closed subspace of a Banach space X and let $y \in X \setminus Y$. Then there exists $x^* \in X'$ such that $x^*(Y) = 0$ but $x^*(y) \neq 0$.*

Next is the Eberlein Smulian theorem which states that a Banach space is reflexive if and only if the closed unit ball is weakly sequentially compact. Actually, only half the theorem is proved here, the more useful only if part. The book by Yoshida [59] has the complete theorem discussed. First here is an interesting lemma for its own sake.

Lemma 21.5.13 *A closed subspace of a reflexive Banach space is reflexive.*

Proof: Let Y be the closed subspace of the reflexive space, X. Consider the following diagram

$$\begin{array}{ccc} Y'' & \overset{i^{**}\ 1\text{-}1}{\rightarrow} & X'' \\ Y' & \overset{i^{*}\ \text{onto}}{\leftarrow} & X' \\ Y & \overset{i}{\rightarrow} & X \end{array}$$

This diagram follows from Theorem 21.2.10 on Page 568, the theorem on adjoints. Now let $y^{**} \in Y''$. Is $y^{**} = J_Y y$ for some $y \in Y$? Since X is reflexive, $i^{**} y^{**} = J_X(y)$ for some y. I want to show that $y \in Y$. If it is not in Y then since Y is closed, there exists $x^* \in X'$ such that $x^*(y) \neq 0$ but $x^*(Y) = 0$. Then $i^* x^* = 0$. Hence

$$0 = y^{**}(i^* x^*) = i^{**} y^{**}(x^*) = J(y)(x^*) = x^*(y) \neq 0,$$

a contradiction. Hence $y \in Y$. Letting J_Y denote the James map from Y to Y'' and $x^* \in X'$,

$$\begin{aligned} y^{**}(i^* x^*) &= i^{**} y^{**}(x^*) = J_X(y)(x^*) \\ &= x^*(y) = x^*(iy) = i^* x^*(y) = J_Y(y)(i^* x^*) \end{aligned}$$

Since i^* is onto, this shows $y^{**} = J_Y(y)$. ■

Theorem 21.5.14 (*Eberlein Smulian*) *The closed unit ball in a reflexive Banach space X, is weakly sequentially compact. By this is meant that if $\{x_n\}$ is contained in the closed unit ball, there exists a subsequence, $\{x_{n_k}\}$ and $x \in X$ such that for all $x^* \in X'$, it follows that $x^*(x_{n_k}) \to x^*(x)$.*

Proof: Let $\{x_n\} \subseteq B \equiv \overline{B(0,1)}$. Let Y be the closure of the linear span of $\{x_n\}$. Thus Y is a separable. It is reflexive because it is a closed subspace of a reflexive space so the above lemma applies. By the Banach Alaoglu theorem, the closed unit ball B_Y^* in Y' is weak $*$ compact. Also by Theorem 21.5.5, B_Y^* is a metric space with a suitable metric. The following diagram illustrates the rest of the argument.

$$\begin{array}{llll} B^{**} & Y'' & \overset{i^{**}\ 1\text{-}1}{\to} & X'' \\ B_Y^* & Y' \text{ weak} * \text{ separable} & \overset{i^*\ \text{onto}}{\leftarrow} & X' \\ B_Y & Y \text{ separable} & \overset{i}{\to} & X \end{array}$$

Thus B_Y^* is complete and totally bounded with respect to this metric and it follows that B_Y^* with the weak $*$ topology is separable. This implies Y' is also separable in the weak $*$ topology. To see this, let $\{y_n^*\} \equiv D$ be a weak $*$ dense set in B_Y^* and let $y^* \in Y'$. Let p be a large enough positive rational number that $y^*/p \in B^*$. Then if A is any finite set from Y, there exists $y_n^* \in D$ such that $\rho_A(y^*/p - y_n^*) < \frac{\varepsilon}{p}$. It follows $p y_n^* \in B_A(y^*, \varepsilon)$ showing that rational multiples of D are weak $*$ dense in Y'. Letting $B_Y = B \cap Y$, this B_Y is the closed unit ball in Y and Y' is weak $*$ separable. Therefore, by Corollary 21.5.10, B_Y is weakly sequentially compact. Thus there exists $\{x_{n_k}\}$ such that $x_{n_k} \to x \in B_Y$ weakly in Y. Letting $x^* \in X^*, i^* x^* \in Y'$ and so

$$x^*(x_{n_k}) = i^* x^*(x_{n_k}) \to i^* x^*(x) = x^*(x)$$

and so in fact, $x_{n_k} \to x$ weakly in X. ■

The following is the form of the Eberlein Smulian theorem which is often used.

Corollary 21.5.15 *Let $\{x_n\}$ be any bounded sequence in a reflexive Banach space X. Then there exists $x \in X$ and a subsequence, $\{x_{n_k}\}$ such that for all $x^* \in X'$, it follows that* $\lim_{k \to \infty} x^*(x_{n_k}) = x^*(x)$.

Proof: If a subsequence, x_{n_k} has $\|x_{n_k}\| \to 0$, then the conclusion follows. Simply let $x = 0$. Suppose then that $\|x_n\|$ is bounded away from 0. That is, $\|x_n\| \in [\delta, C]$. Take a subsequence such that $\|x_{n_k}\| \to a$. Then consider $x_{n_k} / \|x_{n_k}\|$. By the Eberlein Smulian theorem, this subsequence has a further subsequence, $x_{n_{k_j}} / \|x_{n_{k_j}}\|$ which converges weakly to $x \in B$ where B is the closed unit ball. It follows from routine considerations that $x_{n_{k_j}} \to ax$ weakly. ■

21.6 Differential Equations

It is a good idea to do Problems 22-24 at this time. Consider $y' = f(t, y, \lambda), y(t_0) = y_0$ for t near t_0 where $\lambda \in V \subseteq \Lambda$ with V an open subset of Λ some Banach space. Assume $f : (t_0 - \delta, t_0 + \delta) \times U \times V \to Z$ is $C^1((t_0 - \delta, t_0 + \delta) \times U \times V)$ where U is an open subset of Z a Banach space and $u_0 \in U$. Let $\alpha \in (-\delta, \delta)$ and let

$$\alpha s \equiv t - t_0, \phi(s) \equiv y(t) - y_0$$

Thus $\phi(0) = 0$. Also $\phi \in C^1([-1,1]; Z)$ so $\phi \in \mathscr{D}^1 \equiv \{y \in C^1([-1,1], Z) : y(0) = 0\}$. From the above problems, $\mathscr{D}^1$ is a Banach space. It is also the case that

$$\phi'(s) = y'(t)\alpha = \alpha f(\alpha s, y_0 + \phi(s), \lambda)$$

Let L be as in Problem 22, $L\phi = \phi'$. Then the problem reduces to

$$L\phi(s) - \alpha f(\alpha s, y_0 + \phi(s), \lambda) = 0, s \in [-1,1]$$

Let $\mathscr{U}_Z$ be the open of Problem 24, all $u \in \mathscr{D}^1$ (so $u(0) = 0$) such that $u(t) \in U_Z$ an open set in Z containing 0, this for each $t \in [-1,1]$. Let

$$F : (-\delta, \delta) \times U \times \mathscr{U}_Z \times V \to C([-1,1]; Z)$$

be defined by

$$F(\alpha, \tilde{y}_0, \psi, \mu)(s) \equiv L\psi(s) - \alpha f(\alpha s, \tilde{y}_0 + \psi(s), \mu)$$

Are the various partial derivatives continuous?

$$\begin{aligned} & F(\alpha+\beta, \tilde{y}_0, \psi, \mu)(s) - F(\alpha, \tilde{y}_0, \psi, \mu)(s) \\ = \; & \alpha f(\alpha s, \tilde{y}_0 + \psi(s), \mu) - (\alpha+\beta) f((\alpha+\beta)s, \tilde{y}_0 + \psi(s), \mu) \\ = \; & -\beta f(\alpha s, \tilde{y}_0 + \psi(s), \mu) + (\alpha+\beta) \begin{pmatrix} f(\alpha s, \tilde{y}_0 + \psi(s), \mu) \\ -f((\alpha+\beta)s, \tilde{y}_0 + \psi(s), \mu) \end{pmatrix} \\ = \; & -\beta f(\alpha s, \tilde{y}_0 + \psi(s), \mu) - (\alpha+\beta)(D_1 f(\alpha s, \tilde{y}_0 + \psi(s), \mu)\beta s + o(\beta s)) \\ = \; & -\beta f(\alpha s, \tilde{y}_0 + \psi(s), \mu) - \alpha(D_1 f(\alpha s, \tilde{y}_0 + \psi(s), \mu)\beta s + o(\beta s)) \end{aligned}$$

Thus $\alpha \to D_1 F(\alpha, \tilde{y}_0, \psi, \mu)$ is continuous as a map from $(-\delta, \delta)$ to $\mathscr{L}(\mathbb{R}, C([-1,1]; Z))$. Similarly, $\tilde{y}_0 \to D_2 F(\alpha, \tilde{y}_0, \psi, \mu)$ is continuous as a map from U to $\mathscr{L}(Z, C([-1,1]; Z))$. and $\mu \to D_4 F(\alpha, \tilde{y}_0, \psi, \mu)$ is continuous as a map from V to $\mathscr{L}(\Lambda, C([-1,1]; Z))$. What remains is to consider $D_3 F$. Note that $\nu_n \to \nu$ in $\mathscr{D}^1$ implies $\nu_n(s) \to \nu(s)$ in Z for each s.

$$F(\alpha, \tilde{y}_0, \psi + \eta, \mu)(s) - F(\alpha, \tilde{y}_0, \psi, \mu)(s) + L\eta$$

$$\begin{aligned} &= F(\alpha,\tilde{y}_0,\psi(s)+\eta(s),\mu)-F(\alpha,\tilde{y}_0,\psi(s),\mu) \\ &= -\alpha D_3 f(\alpha,\tilde{y}_0,\psi(s),\mu)\eta(s)+o(\eta(s))+L\eta \end{aligned}$$

Now because of the definition of $\mathscr{D}^1$, if $g(\eta)(s) \equiv o(\eta(s))$, then $g(\eta) = o(\eta)$. Indeed, $\frac{\|g(\eta)(s)\|}{\|\eta\|} \leq \frac{\|g(\eta)(s)\|}{\|\eta(s)\|} < \varepsilon$ if $\|\eta\|_{\mathscr{D}^1}$ is small enough. Hence $\|g(\eta)\| \leq \varepsilon\|\eta\|$ if $\|\eta\|$ is small enough. Thus $D_3F(\alpha,\tilde{y}_0,\psi,\mu)(s) = -\alpha D_3 f(\alpha,\tilde{y}_0,\psi,\mu)+L$, so $\psi \to D_3F(\alpha,\tilde{y}_0,\psi,\mu)$ is continuous into $\mathscr{L}(\mathscr{D}^1, C([-1,1];Z))$. Since L is continuous, one to one and onto, the open mapping theorem says that $L^{-1}: C([-1,1];Z) \to \mathscr{D}^1$ is continuous, Problem 23. It follows that for given $y_0 \in U$, $D_3F(0,y_0,0,\lambda) = L$ which is invertible. By the implicit function theorem, there exists a unique $\phi(\alpha,\tilde{y}_0,\mu)$ where ϕ is a C^1 function of $\alpha,\tilde{y}_0,\mu$ for α close enough to 0, $\tilde{y}_0$ close enough to y_0 and μ close enough to λ, say $(\alpha,\tilde{y}_0,\mu) \in (-\sigma,\sigma)\times B(y_0,r)\times B(\lambda,\delta)$ for which

$$F(\alpha,\tilde{y}_0,\phi,\mu)(s) \equiv L\phi(s) - \alpha f(\alpha s,\tilde{y}_0+\phi(s,\tilde{y}_0,\mu),\mu) = 0$$

Pick small positive α. Thus, this α will be fixed. Now go backwards in how this started. Let $t = \alpha s + t_0$ so $t - t_0 \in [-\alpha,\alpha]$. Let $y(t,\tilde{y}_0,\mu) \equiv \phi(s,\tilde{y}_0,\mu)+\tilde{y}_0$ so $y'(t)\alpha = \phi'(s) = L\phi = \alpha f(t,y(t,\tilde{y}_0,\mu),\mu)$ so $y(t_0,\tilde{y}_0,\mu) = y_0, y'(t,y_0,\mu) = f(t,y(t,\tilde{y}_0,\mu),\mu)$.

If f had been C^k instead of just C^1, the same conclusion would follow except now you would have $y(t,\tilde{y}_0,\mu)$ a C^k function of $\tilde{y}_0$ and μ for $(\tilde{y}_0,\mu)$ close to a particular (y_0,λ).

This use of the implicit function theorem to give existence, uniqueness, and differentiable dependence on initial data and a given parameter is extemely significant because it justifies often used procedures for writing a solution to a differential equation in terms of a Taylor series in powers of a parameter.

This proves the following theorem which is an existence and uniqueness theorem for the initial value problem for ordinary differential equations that also gives a description of dependence on the initial data and an arbitrary parameter.

Theorem 21.6.1 *Let $f: (t_0-\delta,t_0+\delta)\times U\times V \to Z$ where U is an open set in Z, V an open set in Λ, some Banach space. Suppose f is*

$$C^k((t_0-\delta,t_0+\delta)\times U\times V, Z).$$

Then if $(y_0,\lambda) \in (U\times V)$, there exists $\alpha > 0$ and a unique solution to

$$y'(t) = f(t,\tilde{y}_0,\mu), y(t_0) = \tilde{y}_0$$

for $t - t_0 \in [-\alpha,\alpha]$ whenever $(\tilde{y}_0,\mu)$ is close enough to (y_0,λ). Denoting this solution as $y(t) = y(t,\tilde{y}_0,\mu)$, it follows that $(\tilde{y}_0,\mu) \to y(t,\tilde{y}_0,\mu)$ is a C^k function.

The case where t is not just real but is allowed to be complex and f is analytic is also available, but I have not discussed analytic functions here. This leads to being able to expand the solution in a power series. See [11] for more on this subject including the analytic case. This case is also in [36].

Example 21.6.2 *Consider $y' = y^2+\varepsilon, y(0) = 0$ where ε is a small real number. Then there is a solution $t \to y(t)$ on a small interval containing 0. According to the above theorem,*

$$y(t) = \sum_{k=0}^{2} a_k(t)\varepsilon^k + o(\varepsilon^2)$$

Use the initial condition to find that $a_k(0)=0$ for each k. Then neglecting higher powers of ε than 2,

$$\sum_{k=0}^{2} a_k'(t)\varepsilon^k = 2\varepsilon^2 a_0(t)a_2(t)+\varepsilon^2 a_1^2(t)+2\varepsilon a_0(t)a_1(t)+a_0^2(t)+\varepsilon$$

Matching the powers of $\varepsilon, a_0'(t)=a_0^2(t), a_0(0)=0$ so $a_0(t)=0$. Similarly $a_1'(t)=1$ so $a_1(t)=t$ and $a_2'(t)=t^2, a_2(0)=0, a_2(t)=\frac{t^3}{3}$. Thus $y(t)=t\varepsilon+\frac{t^3}{3}\varepsilon^2+o(\varepsilon^2)$. The exact solution for positive ε is $\sqrt{\varepsilon}\tan\left(t\sqrt{\varepsilon}\right)$ and this equals $t\varepsilon+\frac{1}{3}t^3\varepsilon^2+o(\varepsilon^2)$. This is an easy example, but the same idea will hold for harder examples for which it might not be possible to find an exact solution.

21.7 Lyapunov Schmidt Procedure

You have $f: X\times\Lambda\to Y$ where here X,Λ are Banach spaces and f is C^p. Suppose $(0,0)\in X\times\Lambda$ and $f(0,0)=0$. Then if $D_1 f(0,0)^{-1}$ is in $\mathscr{L}(Y,X)$, the implicit function theorem says that there exists $x(\lambda)$ a C^p function such that locally $f(x(\lambda),\lambda)=0$. So what if $D_1 f(0,0)$ fails to be one to one? Sometimes this case is also considered. It may be that $D_1 f(0,0)$ is one to one on some subspace and other nice things happen. In particular, suppose the following.

Letting $X_2 \equiv \ker D_1 f(0,0)$ assume

$$X = X_1 \oplus X_2,\ \dim(X_2)<\infty$$

where X_1 is a closed subspace. Thus $D_1 f(0,0)$ is one to one on X_1. We let

$$Y_1 = D_1 f(0,0)(X_1)$$

and suppose that $Y = Y_1\oplus Y_2$ where $\dim(Y_2)<\infty$, and Y_1 is also a closed subspace.

$$X_1 \overset{D_1 f(0,0)}{\to} Y_1 = D_1 f(0,0)(X_1),\ Y_1 \text{ closed} \atop Y = Y_1\oplus Y_2,\ \dim(Y_2)<\infty \quad <\infty$$

By the open mapping theorem, $D_1 f(0,0)^{-1}$ is also continuous.

Let Q be a continuous projection onto Y_1 which is assumed to exist[2] such that $QY_2=0$ and $(I-Q)$ is a projection onto Y_2. Then the equation $f(x(\lambda),\lambda)=0$ can be written as the pair

$$\begin{aligned} Qf(x,\lambda) &= 0 \\ (I-Q)f(x,\lambda) &= 0 \end{aligned}$$

Consider the top. For $x=x_1+x_2$ where $x_i\in X_i$, this is

$$Qf(x_1+x_2,\lambda)=0$$

Then if $g(x_1,x_2,\lambda)=Qf(x_1+x_2,\lambda)$, one has $g: X_1\times X_2\times\Lambda\to Y_1$

$$D_1 g(x_1,x_2,\lambda)h = D_1 Qf(x_1+x_2,\lambda)h,\ h\in X_1.$$

[2] In Hilbert space, the existence of this projection map is obvious and it is assumed that it exists here.

Thus $D_1 g(0,0,0)^{-1}$ is continuous by the open mapping theorem ($D_1 f(0,0)$ is one to one on X_1), and by the implicit function theorem, there is a solution to

$$Qf(x_1 + x_2, \lambda) = 0$$

for $x_1 = x_1(x_2, \lambda)$. (Note how it is important that X_1 and Y_1 be Banach spaces.) Then the other equation yields

$$(I - Q) f(x_1(x_2, \lambda) + x_2, \lambda) = 0$$

and so for fixed λ, this is a finite set of equations of a variable in a finite dimensional space.

This depends on being able to write $X = X_1 \oplus X_2$ where X_1 is closed, $X_2 = \ker D_1 f(0,0)$, a similar situation for $Y = Y_1 \oplus Y_2$. So when does this happen? Are there conditions on $D_1 f(0,0)$ which will cause it to occur?

There are such conditions. For example, $D_1 f(0,0)$ could be a Fredholm operator defined in Definition 21.7.1.

Definition 21.7.1 *Let $T \in \mathscr{L}(X,Y)$. Then this is a Fredholm operator means*

1. $\dim(\ker(T)) < \infty$
2. $\dim(E) < \infty$ *where* $Y = TX \oplus E$

The following are some easy examples in which all that nonsense about things being finite dimensional and part of a direct sum does not need to be considered.

Example 21.7.2 *Say $X = \mathbb{R}^2$ and $\Lambda = \mathbb{R}$. Let $f(x,y,\lambda) = x + xy + y^2 + \lambda$. Then*

$$D_1 f(0,0,0) = (1,0)$$

this 1×2 matrix mapping $\mathbb{R}^2$ to $\mathbb{R}$. Thus $X_2 = (0,\alpha)^T : \alpha \in \mathbb{R}$ and $X_1 = (\alpha,0)^T : \alpha \in \mathbb{R}$. In this case, $Y_1 = \mathbb{R}$ and so $Q = I$. Thus the above reduces to the single equation

$$f((\alpha,0) + (0,\beta), \lambda) = 0$$

and so since $D_1 f(0,0,0)$ is one to one, $x_1 = (\alpha,0) = x_1((0,\beta),\lambda)$. Of course this is completely obvious because if you consider f in the natural way as a function of three variables, then the implicit function theorem immediately gives $x = x(y,\lambda)$ which is essentially the same result. We just write $(\alpha,0)$ in place of α. The first independent variable is a function of the other two.

Example 21.7.3 *Here is another easy example. $\boldsymbol{f} : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}^2$*

$$\boldsymbol{f}(x,y,\lambda) = \begin{pmatrix} x + xy + y^2 + \sin(\lambda) \\ x + y^2 - x^2 + \lambda \end{pmatrix}$$

Then

$$D_1 \boldsymbol{f}(x,y,\lambda) = \begin{pmatrix} 1+y & x+2y \\ 1-2x & 2y \end{pmatrix}$$

So

$$D_1 \boldsymbol{f}((0,0),0) = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$$

Then

$$X_2 = \ker D_1 \boldsymbol{f}((0,0),0) = \left\{ \begin{pmatrix} 0 \\ \beta \end{pmatrix} : \beta \in \mathbb{R} \right\}$$

and $X_1 = \left\{ \begin{pmatrix} \alpha \\ 0 \end{pmatrix} : \alpha \in \mathbb{R} \right\}$ *and clearly* $D_1 \boldsymbol{f}((0,0),0)$ *is indeed one to one on* X_1.

$$D_1 \boldsymbol{f}(\mathbf{0},0)(X_1) = \left\{ \begin{pmatrix} y \\ y \end{pmatrix} : y \in \mathbb{R} \right\} = Y_1$$

In this case, let

$$Q \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \frac{\alpha+\beta}{2} \\ \frac{\alpha+\beta}{2} \end{pmatrix} = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

so $(I-Q) = \begin{pmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{pmatrix}$. *Thus the equations are*

$$\begin{aligned} Q\boldsymbol{f}(\boldsymbol{x},\lambda) &= 0 \\ (I-Q)\boldsymbol{f}(\boldsymbol{x},\lambda) &= 0 \end{aligned}$$

This reduces to

$$\begin{aligned} \begin{pmatrix} -\frac{1}{2}x^2 + \frac{1}{2}xy + x + y^2 + \frac{1}{2}\lambda + \frac{1}{2}\sin\lambda \\ -\frac{1}{2}x^2 + \frac{1}{2}xy + x + y^2 + \frac{1}{2}\lambda + \frac{1}{2}\sin\lambda \end{pmatrix} &= \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\ \begin{pmatrix} \frac{1}{2}x^2 + \frac{1}{2}yx - \frac{1}{2}\lambda + \frac{1}{2}\sin\lambda \\ -\frac{1}{2}x^2 - \frac{1}{2}yx + \frac{1}{2}\lambda - \frac{1}{2}\sin\lambda \end{pmatrix} &= \begin{pmatrix} 0 \\ 0 \end{pmatrix} \end{aligned}$$

Note how in both the top and the bottom, there is only one equation and one can solve for x *in terms of* y,λ *near* $(0,0,0)$ *which is what the above general argument shows. Of course you can see this directly using the implicit function theorem. Then can you solve for* $y = y(\lambda)$? *This would involve trying to solve for* y *as a function of* λ *in the following where* $x(y,\lambda)$ *comes from the first equations.*

$$\frac{1}{2}x^2(y,\lambda) + \frac{1}{2}yx(y,\lambda) - \frac{1}{2}\lambda + \frac{1}{2}\sin\lambda = 0$$

If you can do this, then you would have found (x,y) *as a function of* λ *for small* λ.

In this example, in the top equation, at $(0,0,0)$, $x_y = 0$. Also $x_\lambda = -1$ so $x(y,\lambda) \approx -\lambda$ other than higher order terms for small y,λ. Then in the bottom equation, for all variables very small, you would have $\lambda^2 + y(-\lambda) - \lambda + \sin(\lambda) = 0$, $y(\lambda) = -1 + \frac{\sin(\lambda)}{\lambda} + \lambda$ at least approximately. Thus it seems there is a nonzero solution to the equation $\boldsymbol{f}(x,y,\lambda) = 0$ which is valid for small λ, x, y, this in addition to the zero solution. Note that for small nonzero λ, $-1 + \frac{\sin(\lambda)}{\lambda} + \lambda \neq 0$. It equals approximately $\lambda - \frac{\lambda^2}{3!}$ for small λ from the power series for $\sin$.

In the next example, the same procedure gives a solution to a problem

$$\boldsymbol{f}((x,y),\lambda) = \mathbf{0}$$

such that for small λ, (x,y) is a function of λ which is nonzero and

$$\boldsymbol{f}((0,0),\lambda) = \mathbf{0}$$

Thus for small λ, there are two solutions to the nonlinear system of equations.

Example 21.7.4 *Let*

$$\boldsymbol{f}((x,y),\lambda) = \begin{pmatrix} x+xy+y^2+x\sin(\lambda) \\ x+y^2-x^2+x\lambda \end{pmatrix}$$

In this case $\boldsymbol{f}((0,0),\lambda) = \boldsymbol{0}$ *even though* λ *might not be 0. The Lyapunov Schmidt procedure will be used to show that there are nonzero solutions* $x(\lambda), y(\lambda)$ *such that*

$$f((x(\lambda),y(\lambda)),\lambda) = 0$$

At origin,

$$D_1\boldsymbol{f}((0,0),0) = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$$

Thus $X_1 = \text{span}(e_1)$ and $X_2 = \text{span}(e_2)$. Then $Y_1 = \text{span}(e_1+e_2)$ and $Y_2 = \text{span}(e_1-e_2)$. Also $D_1\boldsymbol{f}((0,0),0)$ is one to one on X_1 and its range is Y_1. Then let

$$Q\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \frac{\alpha+\beta}{2} \\ \frac{\alpha+\beta}{2} \end{pmatrix} = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}\begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

$$(I-Q) = \begin{pmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{pmatrix}$$

Then $Q\boldsymbol{f} = \boldsymbol{0}$ is yields the equation

$$x+\frac{1}{2}x\lambda+\frac{1}{2}x\sin\lambda+\frac{1}{2}xy-\frac{1}{2}x^2+y^2 = 0$$

Also $(I-Q)\boldsymbol{f} = \boldsymbol{0}$ yields the equation

$$\frac{1}{2}x\sin\lambda-\frac{1}{2}x\lambda+\frac{1}{2}xy+\frac{1}{2}x^2 = 0$$

Now consider x_y and x_λ at $(0,0)$ from the first equation. Both of these are easily seen to be 0. Now consider x_{yy}. After some computations, this is seen to be $x_{yy} = -2$. Similarly, $x_{y\lambda}(0,0) = 0, x_{\lambda\lambda}(0,0) = 0$ also. Thus up to terms of degree 3,

$$x(y,\lambda) = -y^2 = \frac{1}{2}(-2)y^2$$

Place this in the bottom equation.

$$\frac{1}{2}y^2\lambda-\frac{1}{2}y^2\sin\lambda-\frac{1}{2}y^3+\frac{1}{2}y^4 = 0$$

Now the idea is to find $y = y(\lambda)$, hopefully nonzero. Divide by y^2 and multiply by 2.

$$y^2-y+\lambda-\sin\lambda = 0$$

Then for small λ this is approximately equal to

$$y^2-y+\frac{\lambda^3}{6} = 0$$

Then a solution for y for small λ is

$$y = \frac{1+\sqrt{1-\frac{2}{3}\lambda^3}}{2}$$

Of course there is another solution as well, when you replace the + with a minus sign. This is the one we want because when $\lambda = 0$ it reduces to $y = 0$. This shows that there exist solutions to the equations $\boldsymbol{f}((x,y),\lambda) = \mathbf{0}$ which for small λ are approximately

$$(x(\lambda), y(\lambda)) = \left(-y^2, \frac{1-\sqrt{1-\frac{2}{3}\lambda^3}}{2}\right)$$

In terms of λ very small,

$$(x(\lambda), y(\lambda)) = \left(\frac{1}{6}\lambda^3 + \frac{1}{6}\sqrt{3}\sqrt{3-2\lambda^3} - \frac{1}{2}, \frac{1-\sqrt{1-\frac{2}{3}\lambda^3}}{2}\right)$$

Using a power series in λ to approximate these functions, this reduces to

$$(x(\lambda), y(\lambda)) = \left(-\frac{1}{36}\lambda^6, \frac{1}{6}\lambda^3 + \frac{1}{36}\lambda^6 + \frac{1}{108}\lambda^9\right)$$

where higher order terms are neglected. Thus there exist other solutions than the zero solution even though λ may be nonzero. Note that in this example, $\boldsymbol{f}((0,0),\lambda) = 0$.

Note that all of this works as well if the function f is defined on an open subset of $X \times \Lambda$ because it is really just an application of the implicit function theorem.

21.8 Exercises

1. Is $\mathbb{N}$ a G_δ set? What about $\mathbb{Q}$? What about a countable dense subset of a complete metric space?

2. $\uparrow$ Let $f : \mathbb{R} \to \mathbb{C}$ be a function. Define the oscillation of a function in $B(x,r)$ by $\omega_r f(x) = \sup\{|f(z) - f(y)| : y,z \in B(x,r)\}$. Define the oscillation of the function at the point, x by $\omega f(x) = \lim_{r\to 0} \omega_r f(x)$. Show f is continuous at x if and only if $\omega f(x) = 0$. Then show the set of points where f is continuous is a G_δ set (try $U_n = \{x : \omega f(x) < \frac{1}{n}\}$). Does there exist a function continuous at only the rational numbers? Does there exist a function continuous at every irrational and discontinuous elsewhere? **Hint:** Suppose D is any countable set, $D = \{d_i\}_{i=1}^\infty$, and define the function, $f_n(x)$ to equal zero for every $x \notin \{d_1, \cdots, d_n\}$ and 2^{-n} for x in this finite set. Then consider $g(x) \equiv \sum_{n=1}^\infty f_n(x)$. Show that this series converges uniformly.

3. Let $f \in C([0,1])$ and suppose $f'(x)$ exists. Show there exists a constant, K, such that $|f(x) - f(y)| \le K|x-y|$ for all $y \in [0,1]$. Let $U_n = \{f \in C([0,1])$ such that for each $x \in [0,1]$ there exists $y \in [0,1]$ such that $|f(x) - f(y)| > n|x-y|\}$. Show that U_n is open and dense in $C([0,1])$ where for $f \in C([0,1])$,

$$\|f\| \equiv \sup\{|f(x)| : x \in [0,1]\}.$$

Show that $\cap_n U_n$ is a dense G_δ set of nowhere differentiable continuous functions. Thus every continuous function is uniformly close to one which is nowhere differentiable.

4. $\uparrow$ Suppose $f(x) = \sum_{k=1}^{\infty} u_k(x)$ where the convergence is uniform and each u_k is a polynomial. Is it reasonable to conclude that $f'(x) = \sum_{k=1}^{\infty} u_k'(x)$? The answer is no. Use Problem 3 and the Weierstrass approximation theorem to show this.

5. Let X be a normed linear space. $A \subseteq X$ is "weakly bounded" if for each $x^* \in X'$, $\sup\{|x^*(x)| : x \in A\} < \infty$, while A is bounded if $\sup\{\|x\| : x \in A\} < \infty$. Show A is weakly bounded if and only if it is bounded.

6. Let f be a 2π periodic locally integrable function on $\mathbb{R}$. The Fourier series for f is given by
$$\sum_{k=-\infty}^{\infty} a_k e^{ikx} \equiv \lim_{n\to\infty} \sum_{k=-n}^{n} a_k e^{ikx} \equiv \lim_{n\to\infty} S_n f(x)$$
where
$$a_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-ikx} f(x)\,dx.$$
Show
$$S_n f(x) = \int_{-\pi}^{\pi} D_n(x-y) f(y)\,dy$$
where
$$D_n(t) = \frac{\sin((n+\frac{1}{2})t)}{2\pi\sin(\frac{t}{2})}.$$
Verify that $\int_{-\pi}^{\pi} D_n(t)\,dt = 1$. Also show that if $g \in L^1(\mathbb{R})$, then
$$\lim_{a\to\infty}\int_{\mathbb{R}} g(x)\sin(ax)\,dx = 0.$$
This last is called the Riemann Lebesgue lemma. **Hint:** For the last part, assume first that $g \in C_c^\infty(\mathbb{R})$ and integrate by parts. Then exploit density of the set of functions in $L^1(\mathbb{R})$.

7. $\uparrow$It turns out that the Fourier series sometimes converges to the function pointwise. Suppose f is 2π periodic and Holder continuous. That is $|f(x)-f(y)| \le K|x-y|^\theta$ where $\theta \in (0,1]$. Show that if f is like this, then the Fourier series converges to f at every point. Next modify your argument to show that if at every point, x, $|f(x+)-f(y)| \le K|x-y|^\theta$ for y close enough to x and larger than x and
$$|f(x-)-f(y)| \le K|x-y|^\theta$$
for every y close enough to x and smaller than x, then $S_n f(x) \to \frac{f(x+)+f(x-)}{2}$, the midpoint of the jump of the function. **Hint:** Use Problem 6.

8. $\uparrow$ Let $Y = \{f$ such that f is continuous, defined on $\mathbb{R}$, and 2π periodic$\}$. Define $\|f\|_Y = \sup\{|f(x)| : x \in [-\pi,\pi]\}$. Show that $(Y, \|\ \|_Y)$ is a Banach space. Let $x \in \mathbb{R}$ and define $L_n(f) = S_n f(x)$. Show $L_n \in Y'$ but $\lim_{n\to\infty}\|L_n\| = \infty$. Show that for each $x \in \mathbb{R}$, there exists a dense G_δ subset of Y such that for f in this set, $|S_n f(x)|$ is

unbounded. Finally, show there is a dense G_δ subset of Y having the property that $|S_n f(x)|$ is unbounded on the rational numbers. **Hint:** To do the first part, let $f(y)$ approximate $\operatorname{sgn}(D_n(x-y))$. Here $\operatorname{sgn} r = 1$ if $r > 0, -1$ if $r < 0$ and 0 if $r = 0$. This rules out one possibility of the uniform boundedness principle. After this, show the countable intersection of dense G_δ sets must also be a dense G_δ set.

9. Let $\alpha \in (0,1]$. Define, for X a compact subset of $\mathbb{R}^p$,

$$C^\alpha(X;\mathbb{R}^n) \equiv \{\boldsymbol{f} \in C(X;\mathbb{R}^n) : \rho_\alpha(\boldsymbol{f}) + \|\boldsymbol{f}\| \equiv \|\boldsymbol{f}\|_\alpha < \infty\}$$

where

$$\|\boldsymbol{f}\| \equiv \sup\{|\boldsymbol{f}(\boldsymbol{x})| : \boldsymbol{x} \in X\}$$

and

$$\rho_\alpha(\boldsymbol{f}) \equiv \sup\{\frac{|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{y})|}{|\boldsymbol{x}-\boldsymbol{y}|^\alpha} : \boldsymbol{x},\boldsymbol{y} \in X,\ \boldsymbol{x} \neq \boldsymbol{y}\}.$$

Show that $(C^\alpha(X;\mathbb{R}^n), \|\cdot\|_\alpha)$ is a complete normed linear space. This is called a Holder space. What would this space consist of if $\alpha > 1$?

10. ↑Let X be the Holder functions which are periodic of period 2π. Define $L_n f(x) = S_n f(x)$ where $L_n : X \to Y$ for Y given in Problem 8. Show $\|L_n\|$ is bounded independent of n. Conclude that $L_n f \to f$ in Y for all $f \in X$. In other words, for the Holder continuous and 2π periodic functions, the Fourier series converges to the function uniformly. **Hint:** $L_n f(x)$ is given by

$$L_n f(x) = \int_{-\pi}^{\pi} D_n(y) f(x-y)\, dy$$

where $f(x-y) = f(x) + g(x,y)$ where $|g(x,y)| \leq C|y|^\alpha$. Use the fact the Dirichlet kernel integrates to one to write

$$\left|\int_{-\pi}^{\pi} D_n(y) f(x-y)\, dy\right| \leq \overbrace{\left|\int_{-\pi}^{\pi} D_n(y) f(x)\, dy\right|}^{=|f(x)|}$$

$$+C\left|\int_{-\pi}^{\pi} \sin\left(\left(n+\frac{1}{2}\right)y\right)\left(g(x,y)/\sin(y/2)\right) dy\right|$$

Show the functions, $y \to g(x,y)/\sin(y/2)$ are bounded in L^1 independent of x and get a uniform bound on $\|L_n\|$. Now use a similar argument to show $\{L_n f\}$ is equicontinuous in addition to being uniformly bounded. In doing this you might proceed as follows. Show

$$\left|L_n f(x) - L_n f(x')\right| \leq \left|\int_{-\pi}^{\pi} D_n(y)\left(f(x-y) - f(x'-y)\right) dy\right|$$

$$\leq \|f\|_\alpha |x-x'|^\alpha + \left|\int_{-\pi}^{\pi} \sin\left(\left(n+\frac{1}{2}\right)y\right)\left(\frac{f(x-y)-f(x)-(f(x'-y)-f(x'))}{\sin\left(\frac{y}{2}\right)}\right) dy\right|$$

Then split this last integral into two cases, one for $|y| < \eta$ and one where $|y| \geq \eta$. If $L_n f$ fails to converge to f uniformly, then there exists $\varepsilon > 0$ and a subsequence, n_k such that $\left\|L_{n_k} f - f\right\|_\infty \geq \varepsilon$ where this is the norm in Y or equivalently the sup norm on $[-\pi, \pi]$. By the Arzela Ascoli theorem, there is a further subsequence, $L_{n_{k_l}} f$ which converges uniformly on $[-\pi, \pi]$. But by Problem 7 $L_n f(x) \to f(x)$.

11. Let X be a normed linear space and let M be a convex open set containing 0. Define

$$\rho(x) = \inf\{t > 0 : \frac{x}{t} \in M\}.$$

Show ρ is a gauge function defined on X. This particular example is called a Minkowski functional. It is of fundamental importance in the study of locally convex topological vector spaces. A set, M, is convex if $\lambda x + (1-\lambda)y \in M$ whenever $\lambda \in [0,1]$ and $x, y \in M$.

12. ↑The Hahn Banach theorem can be used to establish separation theorems. Let M be an open convex set containing 0. Let $x \notin M$. Show there exists $x^* \in X'$ such that $\mathrm{Re}\, x^*(x) \geq 1 > \mathrm{Re}\, x^*(y)$ for all $y \in M$. **Hint:** If $y \in M, \rho(y) < 1$. Show this. If $x \notin M$, $\rho(x) \geq 1$. Try $f(\alpha x) = \alpha \rho(x)$ for $\alpha \in \mathbb{R}$. Then extend f to the whole space using the Hahn Banach theorem and call the result F, show F is continuous, then fix it so F is the real part of $x^* \in X'$.

13. A Banach space is said to be strictly convex if whenever $\|x\| = \|y\|$ and $x \neq y$, then

$$\left\|\frac{x+y}{2}\right\| < \|x\|.$$

$F : X \to X'$ is said to be a duality map if it satisfies the following: a.) $\|F(x)\| = \|x\|$. b.) $F(x)(x) = \|x\|^2$. Show that if X' is strictly convex, then such a duality map exists. The duality map is an attempt to duplicate some of the features of the Riesz map in Hilbert space. This Riesz map is the map which takes a Hilbert space to its dual defined as follows.

$$R(x)(y) = (y, x)$$

The Riesz representation theorem for Hilbert space says this map is onto. **Hint:** For an arbitrary Banach space, let

$$F(x) \equiv \left\{x^* : \|x^*\| \leq \|x\| \text{ and } x^*(x) = \|x\|^2\right\}$$

Show $F(x) \neq \emptyset$ by using the Hahn Banach theorem on $f(\alpha x) = \alpha \|x\|^2$. Next show $F(x)$ is closed and convex. Finally show that you can replace the inequality in the definition of $F(x)$ with an equal sign. Now use strict convexity to show there is only one element in $F(x)$.

14. Prove the following theorem which is an improved version of the open mapping theorem, [14]. Let X and Y be Banach spaces and let $A \in \mathscr{L}(X, Y)$. Then the following are equivalent.

$$AX = Y,$$

$$A \text{ is an open map.}$$

Note this gives the equivalence between A being onto and A being an open map. The open mapping theorem says that if A is onto then it is open.

15. Suppose $D \subseteq X$ and D is dense in X. Suppose $L : D \to Y$ is linear and $\|Lx\| \leq K\|x\|$ for all $x \in D$. Show there is a unique extension of L, $\widetilde{L}$, defined on all of X with $\|\widetilde{L}x\| \leq K\|x\|$ and $\widetilde{L}$ is linear. You do not get uniqueness when you use the Hahn Banach theorem. Therefore, in the situation of this problem, it is better to use this result.

16. $\uparrow$ A Banach space is uniformly convex if whenever $\|x_n\|$, $\|y_n\| \leq 1$ and $\|x_n + y_n\| \to 2$, it follows that $\|x_n - y_n\| \to 0$. Show uniform convexity implies strict convexity (See Problem 13). **Hint:** Suppose it is not strictly convex. Then there exist $\|x\|$ and $\|y\|$ both equal to 1 and $\left\|\frac{x_n+y_n}{2}\right\| = 1$ consider $x_n \equiv x$ and $y_n \equiv y$, and use the conditions for uniform convexity to get a contradiction. It can be shown that L^p is uniformly convex whenever $\infty > p > 1$. See Hewitt and Stromberg [26] or Ray [46]. See Theorem 21.3.10.

17. Show that a closed subspace of a reflexive Banach space is reflexive. This is done in the chapter. However, try to do it yourself.

18. x_n converges weakly to x if for every $x^* \in X'$, $x^*(x_n) \to x^*(x)$. $x_n \rightharpoonup x$ denotes weak convergence. Show that if $\|x_n - x\| \to 0$, then $x_n \rightharpoonup x$.

19. $\uparrow$ Show that if X is uniformly convex, then if $x_n \rightharpoonup x$ and $\|x_n\| \to \|x\|$, it follows $\|x_n - x\| \to 0$. **Hint:** Use Lemma 21.2.9 to obtain $f \in X'$ with $\|f\| = 1$ and $f(x) = \|x\|$. See Problem 16 for the definition of uniform convexity. Now by the weak convergence, you can argue that if $x \neq 0$, $f(x_n/\|x_n\|) \to f(x/\|x\|)$. You also might try to show this in the special case where $\|x_n\| = \|x\| = 1$.

20. Suppose $L \in \mathscr{L}(X,Y)$ and $M \in \mathscr{L}(Y,Z)$. Show $ML \in \mathscr{L}(X,Z)$ and that $(ML)^* = L^*M^*$.

21. This problem gives a simple condition for the subgradient of a convex function to be onto. Let X be a reflexive Banach space and suppose $\phi : X \to (-\infty, \infty]$ is convex, proper, lower semicontinuous. This means

    (a) Convex: $\phi(tx + (1-t)y) \leq t\phi(x) + (1-t)\phi(y)$ for all $t \in [0,1]$.

    (b) Lower semicontinuous: If $x_n \to x$, then $\phi(x) \leq \liminf_{n\to\infty} \phi(x_n)$.

    (c) The subgradient of ϕ at x, denoted as $\partial\phi(x)$ is defined as follows: $y^* \in \partial\phi(x)$ means $y^*(z-x) \leq \phi(z) - \phi(x)$ for any z.

    Suppose then that for all $y^* \in X'$,

    $$\lim_{\|x\|\to\infty} \phi(x) - \langle y^*, x\rangle = \infty$$

    this last condition being called "coercive". Show that under these conditions, you can conclude that $\partial\phi$ is not just nonempty for some x but that in fact every $y^* \in X'$ is contained in some $\partial\phi(x)$. Thus $\partial\phi$ is actually onto. **Hint:** Consider the function $x \to \phi(x) - \langle y^*, x\rangle$. Argue that it is lower semicontinuous. Let

    $$\lambda \equiv \inf\{\phi(x) - \langle y^*, x\rangle : x \in X\}$$

    Let $\{x_n\}$ be a minimizing sequence. Argue that from the coercivity condition, $\|x_n\|$ must be bounded. Now use the Eberlein Smulian theorem, to verify that there is a

weakly convergent subsequence $x_n \to x$ weakly. In finite dimensions, you just use the Heine Borel theorem. You know the epigraph of ϕ intersected with $X \times \mathbb{R}$ is a convex and closed subset of $X \times \mathbb{R}$. Explain why this is so. This will require a separation theorem in infinite dimensional space like Problem 12 above. In finite dimensional space, there isn't much to show here. Next explain why ϕ must be weakly lower semicontinuous. If you can't do this part, just use the theorem that a function which is convex and lower semicontinuous is also weakly lower semicontinuous or specialize to finite dimensions and use advanced calculus. That is, if $x_n \to x$ weakly, then $\phi(x) \leq \liminf_{n\to\infty} \phi(x_n)$. Conclude that $\lambda > -\infty$ and equals $\phi(x) - \langle y^*, x\rangle$ which is no larger than $\phi(z) - \langle y^*, z\rangle$. Now conclude that $y^* \in \partial\phi(x)$.

22. Let Z be a Banach space . Let $\mathscr{D}^1 \equiv \left\{y \in C^1([-1,1],Z) : y(0) = 0\right\}$. Let $\|y\|_{\mathscr{D}^1} \equiv \max(\|y\|_\infty, \|y'\|_\infty)$. Show $\mathscr{D}^1$ is a Banach space.

23. ↑Let $L : \mathscr{D}^1 \to C([-1,1],Z), Ly \equiv y'$. Show L is continuous, defined on $\mathscr{D}^1$ and one to one onto $L\left(\mathscr{D}^1\right)$. Show $L\left(\mathscr{D}^1\right) = C([-1,1],Z)$. Thus L^{-1} is continuous by the open mapping theorem. **Hint:** Adapt the Rieman integral to an integral which has values in a Banach space including the fundamental theorem of calculus. Then if $u \in C([-1,1],Z)$, consider $\int_0^t u(s)\,ds \equiv w(t)$ and argue $Lw = u$.

24. ↑Let U_Z denote an open set in Z. Let $f : U_Z \to Z$ be C^1. Then define for $u \in \mathscr{D}^1$, $f(u)(t) \equiv f(u(t))$. Show that $f\left(\mathscr{D}^1\right) \subseteq C([-1,1],Z)$. If $\mathscr{U}_Z$ consists of $u \in \mathscr{D}^1$ such that $u(t) \in U_Z$ for each $t \in [-1,1]$, show that $\mathscr{U}_Z$ is an open subset of $\mathscr{D}^1$. If $f : U \to Z$ is C^1, show that $f : \mathscr{U}_Z \to \mathscr{D}^1$ is also C^1 and that

$$Df(u)(v)(t) = Df(u(t))(v(t)).$$

Chapter 22

Hilbert Spaces

In this chapter, Hilbert spaces, which have been alluded to earlier are given a complete discussion. These spaces, as noted earlier are just complete inner product spaces.

22.1 Basic Theory

Definition 22.1.1 *Let X be a vector space. An inner product is a mapping from $X \times X$ to $\mathbb{C}$ if X is complex and from $X \times X$ to $\mathbb{R}$ if X is real, denoted by (x,y) which satisfies the following.*

$$(x,x) \geq 0,\ (x,x) = 0 \text{ if and only if } x = 0, \tag{22.1}$$

$$(x,y) = \overline{(y,x)}. \tag{22.2}$$

For $a,b \in \mathbb{C}$ and $x,y,z \in X$,

$$(ax+by,z) = a(x,z) + b(y,z). \tag{22.3}$$

Note that 22.2 and 22.3 imply $(x,ay+bz) = \overline{a}(x,y) + \overline{b}(x,z)$. Such a vector space is called an inner product space.

The Cauchy Schwarz inequality is fundamental for the study of inner product spaces.

Theorem 22.1.2 *(Cauchy Schwarz) In any inner product space*

$$|(x,y)| \leq \|x\|\ \|y\|.$$

Equality holds if and only if x is a multiple of y. The inequality holds under the weaker assumption that $(x,x) \geq 0$ without the stipulation that this happens only if $x = 0$.

Proof: Let $\omega \in \mathbb{C}$, $|\omega| = 1$, and $\overline{\omega}(x,y) = |(x,y)| = \mathrm{Re}(x,y\omega)$. Let $F(t) = (x+ty\omega, x+t\omega y), t \in \mathbb{R}$. If $y = 0$ there is nothing to prove because $(x,0) = (x,0+0) = (x,0)+(x,0)$ and so $(x,0) = 0$. Thus, it can be assumed $y \neq 0$. Then from the axioms of the inner product,

$$F(t) = \|x\|^2 + 2t\,\mathrm{Re}(x,\omega y) + t^2\|y\|^2 \geq 0.$$

This yields $\|x\|^2 + 2t|(x,y)| + t^2\|y\|^2 \geq 0$. Since this inequality holds for all $t \in \mathbb{R}$, it follows from the quadratic formula that $4|(x,y)|^2 - 4\|x\|^2\|y\|^2 \leq 0$. In getting this inequality, it was only necessary to assume $(x,x) \geq 0$.

Consider the claim about equality. Note that if $x = \alpha y$, then equality holds directly. Indeed this is the only way this can happen because if equality holds in the Cauchy Schwarz inequality, $F(t) = 0$ for some real t. This happens only if $x = -t\omega y$ for some t real. ■

Proposition 22.1.3 *For an inner product space, $\|x\| \equiv (x,x)^{1/2}$ does specify a norm.*

Proof: All the axioms are obvious except the triangle inequality. To verify this,

$$\begin{aligned} \|x+y\|^2 &\equiv (x+y,x+y) \equiv \|x\|^2 + \|y\|^2 + 2\,\mathrm{Re}\,(x,y) \\ &\leq \|x\|^2 + \|y\|^2 + 2\,|(x,y)| \\ &\leq \|x\|^2 + \|y\|^2 + 2\,\|x\|\,\|y\| = (\|x\| + \|y\|)^2. \blacksquare \end{aligned}$$

The following lemma is called the parallelogram identity.

Lemma 22.1.4 *In an inner product space,*

$$\|x+y\|^2+\|x-y\|^2=2\|x\|^2+2\|y\|^2.$$

The proof, a straightforward application of the inner product axioms, is left to the reader.

Lemma 22.1.5 *For $x\in H$, an inner product space,*

$$\|x\|=\sup_{\|y\|\leq 1}|(x,y)| \tag{22.4}$$

Proof: By the Cauchy Schwarz inequality, if $x\neq 0$,

$$\|x\|\geq \sup_{\|y\|\leq 1}|(x,y)|\geq \left(x,\frac{x}{\|x\|}\right)=\|x\|.$$

It is obvious that 22.4 holds in the case that $x=0$.

Definition 22.1.6 *A Hilbert space is an inner product space which is complete. Thus a Hilbert space is a Banach space in which the norm comes from an inner product as described above.*

In Hilbert space, one can define a projection map onto closed convex nonempty sets.

Definition 22.1.7 *A set K is convex if whenever $\lambda\in[0,1]$ and $x,y\in K$, $\lambda x+(1-\lambda)y\in K$.*

Theorem 22.1.8 *Let K be a closed convex nonempty subset of a Hilbert space H, and let $x\in H$. Then there exists a unique point $Px\in K$ such that $\|Px-x\|\leq\|y-x\|$ for all $y\in K$.*

Proof: Consider uniqueness. Suppose that z_1 and z_2 are two different elements of K such that for $i=1,2$,

$$\|z_i-x\|\leq\|y-x\| \tag{22.5}$$

for all $y\in K$. Also, note that since K is convex, $\frac{z_1+z_2}{2}\in K$. Therefore, by the parallelogram identity,

$$\begin{aligned}
\|z_1-x\|^2 &\leq \left\|\frac{z_1+z_2}{2}-x\right\|^2=\left\|\frac{z_1-x}{2}+\frac{z_2-x}{2}\right\|^2\\
&= 2\left(\left\|\frac{z_1-x}{2}\right\|^2+\left\|\frac{z_2-x}{2}\right\|^2\right)-\left\|\frac{z_1-z_2}{2}\right\|^2\\
&= \frac{1}{2}\|z_1-x\|^2+\frac{1}{2}\|z_2-x\|^2-\left\|\frac{z_1-z_2}{2}\right\|^2\\
&\leq \|z_1-x\|^2-\left\|\frac{z_1-z_2}{2}\right\|^2,
\end{aligned}$$

where the last inequality holds because of 22.5. Hence $z_1=z_2$ after all and this shows uniqueness.

Now let $\lambda = \inf\{\|x-y\| : y \in K\}$ and let y_n be a minimizing sequence. This means $\{y_n\} \subseteq K$ satisfies $\lim_{n\to\infty} \|x-y_n\| = \lambda$. By the parallelogram identity,

$$\|y_n - x + y_m - x\|^2 + \|y_n - y_m\|^2 = 2\left(\|y_n - x\|^2 + \|y_m - x\|^2\right)$$

and so, since $\|y_n - x + y_m - x\|^2 = 4(\|\frac{y_n+y_m}{2} - x\|^2)$,

$$\begin{aligned} \|y_n - y_m\|^2 &= 2\left(\|y_n - x\|^2 + \|y_m - x\|^2\right) - 4(\|\frac{y_n+y_m}{2} - x\|^2) \\ &\leq 2\left(\|y_n - x\|^2 + \|y_m - x\|^2\right) - 4\lambda^2 \end{aligned}$$

The right side converges as $m,n \to 0$ to 0. Therefore, $\{y_n\}_{n=1}^{\infty}$ is a Cauchy sequence. Since H is complete, $y_n \to y$ for some $y \in H$ which must be in K because K is closed. Therefore

$$\|x-y\| = \lim_{n\to\infty} \|x-y_n\| = \lambda.$$

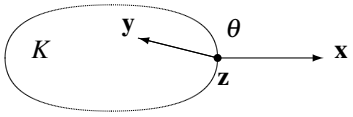
Let $Px = y$. ■

Corollary 22.1.9 *Let K be a closed, convex, nonempty subset of a Hilbert space, H, and let $x \in H$. Then for $z \in K$, $z = Px$ if and only if*

$$\mathrm{Re}(x-z, y-z) \leq 0 \tag{22.6}$$

for all $y \in K$.

Before proving this, consider what it says in the case where the Hilbert space is $\mathbb{R}^n$.



Condition 22.6 says the angle θ, shown in the diagram, is always obtuse. Remember from calculus, the sign of $\boldsymbol{x} \cdot \boldsymbol{y}$ is the same as the sign of the cosine of the included angle between $\boldsymbol{x}$ and $\boldsymbol{y}$. Thus, in finite dimensions, the conclusion of this corollary says that $z = Px$ exactly when the angle of the indicated angle is obtuse. Surely the picture suggests this is reasonable.

The inequality 22.6 is an example of a variational inequality and this corollary characterizes the projection of x onto K as the solution of this variational inequality.

Proof of Corollary: Let $z \in K$ and let $y \in K$ also. Since K is convex, it follows that if $t \in [0,1]$,

$$z + t(y-z) = (1-t)z + ty \in K.$$

Furthermore, every point of K can be written in this way. (Let $t = 1$ and $y \in K$.) Therefore, $z = Px$ if and only if for all $y \in K$ and $t \in [0,1]$,

$$\|x - (z + t(y-z))\|^2 = \|(x-z) - t(y-z)\|^2 \geq \|x-z\|^2$$

for all $t \in [0,1]$ and $y \in K$ if and only if for all $t \in [0,1]$ and $y \in K$

$$\|x-z\|^2 + t^2\|y-z\|^2 - 2t\,\mathrm{Re}\,(x-z, y-z) \geq \|x-z\|^2$$

If and only if for all $t \in [0,1]$,

$$t^2 \|y-z\|^2 - 2t \operatorname{Re}(x-z, y-z) \geq 0. \tag{22.7}$$

Now this is equivalent to 22.7 holding for all $t \in (0,1)$. Therefore, dividing by $t \in (0,1)$, 22.7 is equivalent to

$$t \|y-z\|^2 - 2 \operatorname{Re}(x-z, y-z) \geq 0$$

for all $t \in (0,1)$ which is equivalent to 22.6. ■

Corollary 22.1.10 *Let K be a nonempty convex closed subset of a Hilbert space, H. Then the projection map P is continuous. In fact, $|Px - Py| \leq |x-y|$.*

Proof: Let $x, x' \in H$. Then by Corollary 22.1.9,

$$\operatorname{Re}(x' - Px', Px - Px') \leq 0, \ \operatorname{Re}(x - Px, Px' - Px) \leq 0$$

Hence

$$\begin{aligned} 0 &\leq \operatorname{Re}(x - Px, Px - Px') - \operatorname{Re}(x' - Px', Px - Px') \\ &= \operatorname{Re}(x - x', Px - Px') - |Px - Px'|^2 \end{aligned}$$

and so $|Px - Px'|^2 \leq |x - x'| |Px - Px'|$. ■

The next corollary is a more general form for the Brouwer fixed point theorem.

Corollary 22.1.11 *Let $\boldsymbol{f} : K \to K$ where K is a convex compact subset of $\mathbb{R}^n$. Then $\boldsymbol{f}$ has a fixed point.*

Proof: Let $K \subseteq \overline{B(\mathbf{0}, R)}$ and let P be the projection map onto K. Then consider the map $\boldsymbol{f} \circ P$ which maps $\overline{B(\mathbf{0}, R)}$ to $\overline{B(\mathbf{0}, R)}$ and is continuous. By the Brouwer fixed point theorem for balls, this map has a fixed point. Thus there exists $\boldsymbol{x}$ such that $(\boldsymbol{f} \circ P)(\boldsymbol{x}) = \boldsymbol{x}$. Now the equation also requires $\boldsymbol{x} \in K$ and so $P(\boldsymbol{x}) = \boldsymbol{x}$. Hence $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{x}$. ■

Recall the following definition from linear algebra about direct sum notation.

Definition 22.1.12 *Let H be a vector space and let U and V be subspaces. $U \oplus V = H$ if every element of H can be written as a sum of an element of U and an element of V in a unique way.*

The case where the closed convex set is a closed subspace is of special importance and in this case the above corollary implies the following.

Corollary 22.1.13 *Let K be a closed subspace of a Hilbert space H, and let $x \in H$. Then for $z \in K$, $z = Px$ if and only if*

$$(x - z, y) = 0 \tag{22.8}$$

for all $y \in K$. Furthermore, $H = K \oplus K^{\perp}$ where

$$K^{\perp} \equiv \{x \in H : (x,k) = 0 \text{ for all } k \in K\}$$

and

$$\|x\|^2 = \|x - Px\|^2 + \|Px\|^2. \tag{22.9}$$

Proof: Since K is a subspace, the condition 22.6 implies $\mathrm{Re}(x-z,y)\leq 0$ for all $y\in K$. Replacing y with $-y$, it follows $\mathrm{Re}(x-z,-y)\leq 0$ which implies $\mathrm{Re}(x-z,y)\geq 0$ for all y. Therefore, $\mathrm{Re}(x-z,y)=0$ for all $y\in K$. Now let $|\alpha|=1$ and $\alpha(x-z,y)=|(x-z,y)|$. Since K is a subspace, it follows $\overline{\alpha}y\in K$ for all $y\in K$. Therefore,

$$0=\mathrm{Re}(x-z,\overline{\alpha}y)=(x-z,\overline{\alpha}y)=\alpha(x-z,y)=|(x-z,y)|.$$

This shows that $z=Px$, if and only if 22.8.

For $x\in H$, $x=x-Px+Px$ and from what was just shown, $x-Px\in K^{\perp}$ and $Px\in K$. This shows that $K^{\perp}+K=H$. Is there only one way to write a given element of H as a sum of a vector in K with a vector in $K^{\perp}$? Suppose $y+z=y_1+z_1$ where $z,z_1\in K^{\perp}$ and $y,y_1\in K$. Then $(y-y_1)=(z_1-z)$ and so from what was just shown, $(y-y_1,y-y_1)=(y-y_1,z_1-z)=0$ which shows $y_1=y$ and consequently $z_1=z$. Finally, letting $z=Px$,

$$\begin{aligned}\|x\|^2 &= (x-z+z,x-z+z)=\|x-z\|^2+(x-z,z)+(z,x-z)+\|z\|^2\\ &= \|x-z\|^2+\|z\|^2\ \blacksquare\end{aligned}$$

The following theorem is called the Riesz representation theorem for the dual of a Hilbert space. If $z\in H$ then define an element $f\in H'$ by the rule $(x,z)\equiv f(x)$. It follows from the Cauchy Schwarz inequality and the properties of the inner product that $f\in H'$. The Riesz representation theorem says that all elements of H' are of this form.

Theorem 22.1.14 *Let H be a Hilbert space and let $f\in H'$. Then there exists a unique $z\in H$ such that $f(x)=(x,z)$ for all $x\in H$.*

Proof: Letting $y,w\in H$ the assumption that f is linear implies

$$f(yf(w)-f(y)w)=f(w)f(y)-f(y)f(w)=0$$

which shows that $yf(w)-f(y)w\in f^{-1}(0)$, which is a closed subspace of H since f is continuous. If $f^{-1}(0)=H$, then f is the zero map and $z=0$ is the unique element of H which satisfies $f(x)=(x,z)$. If $f^{-1}(0)\neq H$, pick $u\notin f^{-1}(0)$ and let $w\equiv u-Pu\neq 0$. Thus Corollary 22.1.13 implies $(y,w)=0$ for all $y\in f^{-1}(0)$. In particular, let $y=xf(w)-f(x)w$ where $x\in H$ is arbitrary. Therefore,

$$0=(f(w)x-f(x)w,w)=f(w)(x,w)-f(x)\|w\|^2.$$

Thus, solving for $f(x)$ and using the properties of the inner product,

$$f(x)=\left(x,\frac{\overline{f(w)}w}{\|w\|^2}\right)$$

Let $z=\overline{f(w)}w/\|w\|^2$. This proves the existence of z. If $f(x)=(x,z_i)$ $i=1,2$, for all $x\in H$, then for all $x\in H$, then $(x,z_1-z_2)=0$ which implies, upon taking $x=z_1-z_2$ that $z_1=z_2$. $\blacksquare$

If $R:H\to H'$ is defined by $Rx(y)\equiv(y,x)$, the Riesz representation theorem above states this map is onto. This map is called the Riesz map. It is routine to show R is conjugate linear and $\|Rx\|=\|x\|$. In fact,

$$\begin{aligned}R(\alpha x+\beta y)(u) &\equiv (u,\alpha x+\beta y)=\bar{\alpha}(u,x)+\bar{\beta}(u,y)\\ &\equiv \bar{\alpha}Rx(u)+\bar{\beta}Ry(u)=\left(\bar{\alpha}Rx+\bar{\beta}Ry\right)(u)\end{aligned}$$

so it is conjugate linear meaning it goes across plus signs and you factor out conjugates.

$$\|Rx\|\equiv\sup_{\|y\|\leq 1}|Rx(y)|\equiv\sup_{\|y\|\leq 1}|(y,x)|=\|x\|$$

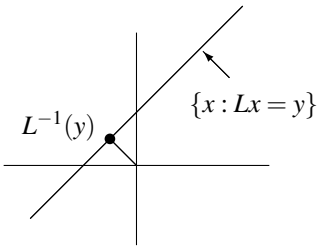
22.2 The Hilbert Space $L(U)$

Let $L \in \mathscr{L}(U,H)$. Then one can consider the image of $L, L(U)$ as a Hilbert space. This is another interesting application of Theorem 22.1.8. First here is a definition which involves abominable and atrociously misleading notation which nevertheless seems to be well accepted.

Definition 22.2.1 *Let $L \in \mathscr{L}(U,H)$, the bounded linear maps from U to H for U, H Hilbert spaces. For $y \in L(U)$, let $L^{-1}y$ denote the unique vector in*

$$\{x \in U : Lx = y\} \equiv M_y$$

which is closest in U to 0.



Note this is a good definition because $\{x \in U : Lx = y\}$ is closed thanks to the continuity of L and it is obviously convex. Thus Theorem 22.1.8 applies. With this definition define an inner product on $L(U)$ as follows. For $y, z \in L(U)$,

$$(y,z)_{L(U)} \equiv \left(L^{-1}y, L^{-1}z\right)_U$$

The notation is abominable because $L^{-1}(y)$ is the normal notation for M_y.

In terms of linear algebra, this L^{-1} is the Moore Penrose inverse. There you obtain the least squares solution x to $Lx = y$ which has smallest norm. Here there is an actual solution and among those solutions you get the one which has least norm. Of course a real honest solution is also a least squares solution so this is the Moore Penrose inverse restricted to $L(U)$.

Lemma 22.2.2 *In the context of the above definition, $L^{-1}(y)$ is characterized by*

$$\begin{aligned} \left(L^{-1}(y), x\right)_U &= 0 \text{ for all } x \in \ker(L) \\ L\left(L^{-1}(y)\right) &= y, \ \left(L^{-1}(y) \in M_y\right) \end{aligned}$$

In addition to this, L^{-1} is linear and the above definition does define an inner product.

Proof: By definition, $L^{-1}(y)$ is the unique point of the closed convex set M_y which is closest to 0 in U. Thus it is characterized by

$$\left(0 - L^{-1}(y), u - L^{-1}(y)\right)_U \leq 0$$

for all $u \in M_y$. Note that $L\left(u - L^{-1}(y)\right) = y - y = 0$. Also, if $v \in \ker(L)$, then if $u = L^{-1}(y) + v$, then $u - L^{-1}(y) \in \ker(L)$. Thus a generic element of $\ker(L)$ is $u - L^{-1}(y)$

for $u \in M_y$ and $L^{-1}(y)$ is therefore characterized by $\left(L^{-1}(y), v\right)_U = 0$ for all $v \in \ker(L)$ because $\ker(L)$ is a subspace. Also, $L\left(L^{-1}(y)\right) = y$. Now from this characterization of L^{-1}, it is obvious that L^{-1} is linear. The inner product is well defined because $L^{-1}(y)$ is uniquely determined. Does it satisfy the axioms? Say $0 = (y,y)_{L(U)}$. Then $L^{-1}(y) = 0$ and so doing L to both sides, $y = 0$. It is clear that $(y,z)_{L(U)} = \overline{(z,y)}_{L(U)}$ because these are defined in terms of a given inner product on U.

$$\begin{aligned}(ay+b\hat{y},z)_{L(U)} &\equiv \left(L^{-1}(ay+b\hat{y}), L^{-1}z\right)_U \\ &= \left(aL^{-1}(y)+bL^{-1}(\hat{y}), L^{-1}z\right)_U\end{aligned}$$

$$\begin{aligned}&= a\left(L^{-1}(y), L^{-1}z\right)_U + b\left(L^{-1}(\hat{y}), L^{-1}z\right)_U \\ &= a(y,z)_{L(U)} + b(\hat{y},z)_{L(U)}\end{aligned}$$

Thus this is an inner product as claimed. ■

With the above definition, here is the main result.

Theorem 22.2.3 *Let U, H be Hilbert spaces and let $L \in \mathscr{L}(U,H)$. Then Definition 22.2.1 makes $L(U)$ into a Hilbert space. Also $L : U \to L(U)$ is continuous and $L^{-1} : L(U) \to U$ is continuous. Also,*

$$\|L\|_{\mathscr{L}(U,H)} \|Lx\|_{L(U)} \geq \|Lx\|_H \tag{22.10}$$

If U is separable, so is $L(U)$. Also $\left(L^{-1}(y), x\right) = 0$ for all $x \in \ker(L)$, and $L^{-1} : L(U) \to U$ is linear. Also, in case that L is one to one, both L and L^{-1} preserve norms.

Proof: First consider the claim that $L : U \to L(U)$ is continuous and $L^{-1} : L(U) \to U$ is also continuous.

$$\|Lu\|^2_{L(U)} = \left(L^{-1}(Lu), L^{-1}(Lu)\right)_U \leq \|u\|^2_U$$

(Recall that $L^{-1}(Lu)$ is the smallest vector in U which maps to Lu. Since u is mapped by L to Lu, it follows that $\left\|L^{-1}(L(u))\right\|_U \leq \|u\|_U$.) Hence L is continuous.

Next, why is L^{-1} continuous? By definition of the norm, $\left\|L^{-1}(y)\right\|^2_U \equiv \|y\|^2_{L(U)}$. Thus L^{-1} is continuous and $\left\|L^{-1}\right\|_{\mathscr{L}(L(U),U)} = 1$.

Why is $L(U)$ a Hilbert space? Let $\{y_n\}$ be a Cauchy sequence in $L(U)$.

$$\|y_n - y_m\|^2_{L(U)} \equiv \left\|L^{-1}(y_n - y_m)\right\|^2_U$$

Then from what was just observed, it follows that $L^{-1}(y_n)$ is a Cauchy sequence in U. Hence $L^{-1}(y_n) \to x \in U$. Then by continuity of L just shown, $y_n \to Lx$. This shows that $L(U)$ is a Hilbert space. It was already shown that it is an inner product space and this has shown that it is complete.

If $x \in U$, then $\|Lx\|_H \leq \|L\|_{\mathscr{L}(U,H)} \|x\|_U$. It follows that

$$\begin{aligned}\|L(x)\|_H &= \left\|L\left(L^{-1}(L(x))\right)\right\|_H \leq \|L\|_{\mathscr{L}(U,H)} \left\|L^{-1}(L(x))\right\|_U \\ &= \|L\|_{\mathscr{L}(U,H)} \|L(x)\|_{L(U)}.\end{aligned}$$

This verifies 22.10.

If U is separable, then letting D be a countable dense subset, it follows from the continuity of the operators L, L^{-1} discussed above that $L(D)$ is separable in $L(U)$. To see this, note that

$$\begin{aligned} \|Lx_n - Lx\|_{L(U)} &= \left\|L\left(L^{-1}(Lx_n - Lx)\right)\right\| \leq \|L\|_{\mathscr{L}(U,H)} \left\|L^{-1}(L(x_n - x))\right\|_U \\ &\leq \|L\|_{\mathscr{L}(U,H)} \|x_n - x\|_U \end{aligned}$$

As before, $L^{-1}(L(x_n - x))$ is the smallest vector which maps onto $L(x_n - x)$ and so its norm is no larger than $\|x_n - x\|_U$.

Consider the last claim. If L is one to one, then for $y \in L(U)$, there is only one vector which maps to y. Therefore, $L^{-1}(L(x)) = x$.Hence for $y \in L(U)$, $\|y\|_{L(U)} \equiv \left\|L^{-1}(y)\right\|_U$. Also,

$$\|Lu\|_{L(U)} \equiv \left\|L^{-1}(L(u))\right\|_U \equiv \|u\|_U$$

Thus when L is one to one, $\|L\|_{\mathscr{L}(U,L(U))} = 1$. ■

22.3 Approximations in Hilbert Space

The Gram Schmidt process applies in any vector space which has an inner product.

Theorem 22.3.1 *Let $\{x_1, \cdots, x_n\}$ be a basis for M a subspace of H a Hilbert space. Then there exists an orthonormal basis for M, $\{u_1, \cdots, u_n\}$ which has the property that for each $k \leq n$, span$(x_1, \cdots, x_k) =$ span$(u_1, \cdots, u_k)$. Also if $\{x_1, \cdots, x_n\} \subseteq H$, then the finite dimensional subspace* span$(x_1, \cdots, x_n)$ *is a closed subspace.*

Proof: Let $\{x_1, \cdots, x_n\}$ be a basis for M. Let $u_1 \equiv x_1 / |x_1|$. Thus for $k = 1$, span$(u_1) =$ span(x_1) and $\{u_1\}$ is an orthonormal set. Now suppose for some $k < n$, $u_1, \cdots, u_k$ have been chosen such that $(u_j \cdot u_l) = \delta_{jl}$ and span$(x_1, \cdots, x_k) =$ span$(u_1, \cdots, u_k)$. Then define

$$u_{k+1} \equiv \frac{x_{k+1} - \sum_{j=1}^{k} (x_{k+1} \cdot u_j) u_j}{\left| x_{k+1} - \sum_{j=1}^{k} (x_{k+1} \cdot u_j) u_j \right|}, \tag{22.11}$$

where the denominator is not equal to zero because the x_j form a basis and so

$$x_{k+1} \notin \text{span}(x_1, \cdots, x_k) = \text{span}(u_1, \cdots, u_k)$$

Thus by induction,

$$u_{k+1} \in \text{span}(u_1, \cdots, u_k, x_{k+1}) = \text{span}(x_1, \cdots, x_k, x_{k+1}).$$

Also, $x_{k+1} \in$ span$(u_1, \cdots, u_k, u_{k+1})$ which is seen easily by solving 22.11 for x_{k+1} and it follows

$$\text{span}(x_1, \cdots, x_k, x_{k+1}) = \text{span}(u_1, \cdots, u_k, u_{k+1}).$$

If $l \leq k$,

$$(u_{k+1} \cdot u_l) = C\left((x_{k+1} \cdot u_l) - \sum_{j=1}^{k} (x_{k+1} \cdot u_j)(u_j \cdot u_l)\right)$$

$$= C\left((x_{k+1} \cdot u_l) - \sum_{j=1}^{k} (x_{k+1} \cdot u_j)\delta_{lj}\right) = C((x_{k+1} \cdot u_l) - (x_{k+1} \cdot u_l)) = 0.$$

The vectors, $\{u_j\}_{j=1}^n$, generated in this way are therefore an orthonormal basis because each vector has unit length.

Consider the second claim about finite dimensional subspaces. Without loss of generality, assume $\{x_1,\cdots,x_n\}$ is linearly independent. If it is not, delete vectors until a linearly independent set is obtained. Then by the first part,

$$\text{span}(x_1,\cdots,x_n) = \text{span}(u_1,\cdots,u_n) \equiv M$$

where the u_i are an orthonormal set of vectors. Suppose $\{y_k\} \subseteq M$ and $y_k \to y \in H$. Is $y \in M$? Let $y_k \equiv \sum_{j=1}^n c_j^k u_j$. Then let $\boldsymbol{c}^k \equiv \left(c_1^k,\cdots,c_n^k\right)^T$. Then

$$\begin{aligned}\left|\boldsymbol{c}^k - \boldsymbol{c}^l\right|^2 &\equiv \sum_{j=1}^n \left|c_j^k - c_j^l\right|^2 = \left(\sum_{j=1}^n \left(c_j^k - c_j^l\right) u_j, \sum_{j=1}^n \left(c_j^k - c_j^l\right) u_j\right) \\ &= \|y_k - y_l\|^2\end{aligned}$$

which shows $\left\{\boldsymbol{c}^k\right\}$ is a Cauchy sequence in $\mathbb{F}^n$ and so it converges to $\boldsymbol{c} \in \mathbb{F}^n$. Thus

$$y = \lim_{k\to\infty} y_k = \lim_{k\to\infty} \sum_{j=1}^n c_j^k u_j = \sum_{j=1}^n c_j u_j \in M. \blacksquare$$

Theorem 22.3.2 *Let M be the span of $\{u_1,\cdots,u_n\}$ in a Hilbert space H and let $y \in H$. Then Py is given by*

$$Py = \sum_{k=1}^n (y,u_k)\, u_k \tag{22.12}$$

and the distance is given by

$$\sqrt{|y|^2 - \sum_{k=1}^n |(y,u_k)|^2}. \tag{22.13}$$

Proof:

$$\begin{aligned}\left(y - \sum_{k=1}^n (y,u_k)\, u_k, u_p\right) &= (y,u_p) - \sum_{k=1}^n (y,u_k)\,(u_k,u_p) \\ &= (y,u_p) - (y,u_p) = 0\end{aligned}$$

It follows that $(y - \sum_{k=1}^n (y,u_k)\, u_k, u) = 0$ for all $u \in M$ and so by Corollary 22.1.13 this verifies 22.12.

The square of the distance, d is given by

$$\begin{aligned}d^2 &= \left(y - \sum_{k=1}^n (y,u_k)\, u_k, y - \sum_{k=1}^n (y,u_k)\, u_k\right) \\ &= |y|^2 - 2\sum_{k=1}^n |(y,u_k)|^2 + \sum_{k=1}^n |(y,u_k)|^2\end{aligned}$$

and this shows 22.13. $\blacksquare$

22.4 Orthonormal Sets

The concept of an orthonormal set of vectors is a generalization of the notion of the standard basis vectors of $\mathbb{R}^n$ or $\mathbb{C}^n$.

Definition 22.4.1 *Let H be a Hilbert space. $S \subseteq H$ is called an orthonormal set if $\|x\| = 1$ for all $x \in S$ and $(x,y) = 0$ if $x, y \in S$ and $x \neq y$. For any set, D,*

$$D^{\perp} \equiv \{x \in H : (x,d) = 0 \text{ for all } d \in D\}.$$

If S is a set, $\operatorname{span}(S)$ is the set of all finite linear combinations of vectors from S.

You should verify that $D^{\perp}$ is always a closed subspace of H. It is assumed that our Hilbert spaces are not $\{0\}$.

Theorem 22.4.2 *In any separable Hilbert space H, there exists a countable orthonormal set, $S = \{x_i\}$ such that the span of these vectors is dense in H. Furthermore, if $\operatorname{span}(S)$ is dense, then for $x \in H$,*

$$x = \sum_{i=1}^{\infty} (x, x_i) x_i \equiv \lim_{n \to \infty} \sum_{i=1}^{n} (x, x_i) x_i. \tag{22.14}$$

Also, $(x,y) = \sum_{i=1}^{\infty} (x, x_i) \overline{(y, y_i)}$.

Proof: Let $\mathscr{F}$ denote the collection of all orthonormal subsets of H. $\mathscr{F}$ is nonempty because some $\{x\} \in \mathscr{F}$ where $\|x\| = 1$. The set, $\mathscr{F}$ is a partially ordered set with the order given by set inclusion. By the Hausdorff maximal theorem, there exists a maximal chain, $\mathfrak{C}$ in $\mathscr{F}$. Then let $S \equiv \cup \mathfrak{C}$. It follows S must be a maximal orthonormal set of vectors. This is because if $x, y \in S$, then both vectors are in a single $C \in \mathfrak{C}$. Therefore, $(x,y) = 0$ or one.

It remains to verify that S is countable $\operatorname{span}(S)$ is dense, and the condition, 22.14 holds. To see S is countable note that if $x, y \in S$, then

$$\|x - y\|^2 = \|x\|^2 + \|y\|^2 - 2\operatorname{Re}(x,y) = \|x\|^2 + \|y\|^2 = 2.$$

Therefore, the open sets, $B\left(x, \frac{1}{2}\right)$ for $x \in S$ are disjoint and cover S. Since H is assumed to be separable, there exists a point from a countable dense set in each of these disjoint balls showing there can only be countably many of the balls and that consequently, S is countable as claimed.

It remains to verify 22.14 and that $\operatorname{span}(S)$ is dense. If $\operatorname{span}(S)$ is not dense, then $\overline{\operatorname{span}(S)}$ is a closed proper subspace of H and letting $y \notin \overline{\operatorname{span}(S)}, z \equiv \frac{y - Py}{\|y - Py\|} \in \operatorname{span}(S)^{\perp}$ where P is the projection map mentioned earlier. But then $S \cup \{z\}$ would be a larger orthonormal set of vectors contradicting the maximality of S.

It remains to verify 22.14. Let $S = \{x_i\}_{i=1}^{\infty}$ and consider the problem of choosing the constants, c_k in such a way as to minimize the expression

$$\left\| x - \sum_{k=1}^{n} c_k x_k \right\|^2 =$$

$$\|x\|^2 + \sum_{k=1}^{n} |c_k|^2 - \sum_{k=1}^{n} \overline{c_k} (x, x_k) - \sum_{k=1}^{n} c_k \overline{(x, x_k)}.$$

This equals $\|x\|^2 + \sum_{k=1}^n |c_k - (x,x_k)|^2 - \sum_{k=1}^n |(x,x_k)|^2$ and therefore, this minimum is achieved when $c_k = (x,x_k)$ and equals $\|x\|^2 - \sum_{k=1}^n |(x,x_k)|^2$. Now since span (S) is dense, if n large enough then for some choice of constants c_k, $\|x - \sum_{k=1}^n c_k x_k\|^2 < \varepsilon$. However, from what was just shown,

$$\left\| x - \sum_{i=1}^n (x,x_i)\, x_i \right\|^2 \leq \left\| x - \sum_{k=1}^n c_k x_k \right\|^2 < \varepsilon$$

showing that $\lim_{n\to\infty} \sum_{i=1}^n (x,x_i)\, x_i = x$ as claimed.

For the last claim, from what was just shown and the observation that if $x_n \to x$ and $y_n \to y$, then $(x_n, y_n) \to (x,y)$,

$$\begin{aligned} (x,y) &= \lim_{n\to\infty} \left(\sum_{i=1}^n (x,x_i)\, x_i, \sum_{j=1}^n (y,x_j)\, x_j \right) = \lim_{n\to\infty} \sum_{i,j\leq n} (x,x_i)\, \overline{(y,x_j)}\, (x_i,x_j) \\ &= \lim_{n\to\infty} \sum_{i=1}^n (x,x_i)\, \overline{(y,x_i)} = \sum_{i=1}^\infty (x,x_i)\, \overline{(y,y_i)} \blacksquare \end{aligned}$$

The proof of this theorem contains the following corollary.

Corollary 22.4.3 *Let S be any orthonormal set of vectors and let*

$$\{x_1, \cdots, x_n\} \subseteq S.$$

Then if $x \in H$

$$\left\| x - \sum_{k=1}^n c_k x_k \right\|^2 \geq \left\| x - \sum_{i=1}^n (x,x_i)\, x_i \right\|^2$$

for all choices of constants, c_k. In addition to this, Bessel's inequality follows,

$$\|x\|^2 \geq \sum_{k=1}^n |(x,x_k)|^2 .$$

If S is countable and span (S) *is dense, then letting $\{x_i\}_{i=1}^\infty = S$, 22.14 follows.*

Corollary 22.4.4 *If V is a closed subspace of a Hilbert space H and if*

$$V = \overline{\text{span}\left(\{u_k\}_{k=1}^N \right)}$$

where the u_k are an orthonormal set and $N \leq \infty$, then for P the projection map onto V, it follows that $Py = \sum_{k=1}^N (y,u_k)\, u_k$ and when $N = \infty$ the series converges to Py. In particular, when $y \in V$, $y = Py = \sum_{k=1}^N (y,u_k)\, u_k$.

Proof: The case where $N < \infty$ was done above. For $y \in H$,

$$\left(y - \sum_{k=1}^\infty (y,u_k)\, u_k, u_m \right) = (y,u_m) - (y,u_m) = 0$$

assuming the series makes sense. Therefore, if this happens, then $Py = \sum_{k=1}^\infty (y,u_k)\, u_k$. Now note that $(y,u_k) = (y - Py, u_k) + (Py,u_k) = (Py,u_k)$. Now, by assumption, there are scalars

c_k^n such that $\lim_{n\to\infty}\left\|Py-\sum_{k=1}^n c_k^n u_k\right\| = 0$. Then by Corollary 22.4.3 and what was just observed,

$$0 = \lim_{n\to\infty}\left\|Py-\sum_{k=1}^n (Py,u_k)\,u_k\right\| = \lim_{n\to\infty}\left\|Py-\sum_{k=1}^n (y,u_k)\,u_k\right\|$$

and so the sum converges to Py as claimed. ■

22.5 Compact Operators in Hilbert Space

Definition 22.5.1 *Let $A\in\mathscr{L}(H,H)$ where H is a Hilbert space. So the map, $x\to(Ax,y)$ is continuous and linear. By the Riesz representation theorem, there exists a unique element of H, denoted by A^*y such that*

$$(Ax,y) = (x,A^*y).$$

It is clear $y\to A^*y$ is linear and continuous. It is linear because

$$\left(\begin{array}{c}(x,A^*(ay+bz)) = \bar{a}(Ax,y)+\bar{b}(Ax,z)\\ = \bar{a}(x,A^*y)+\bar{b}(x,A^*z) = (x,aA^*y+bA^*z)\end{array}\right)$$

A^* is called the adjoint of A. A is a self adjoint operator if $A=A^*$. Thus for a self adjoint operator, $(Ax,y)=(x,Ay)$ for all $x,y\in H$ and so $(Ax,x)=\overline{(x,Ax)}=\overline{(Ax,x)}$ so (Ax,x) is real. A is a compact operator if whenever $\{x_k\}$ is a bounded sequence, there exists a convergent subsequence of $\{Ax_k\}$. Equivalently, A maps bounded sets to sets whose closures are compact.

There is an important observation about the range of a compact operator. It is a general result so I will express it in terms of Banach space.

Proposition 22.5.2 *Let X be a Banach space and let $A\in\mathscr{L}(X,X)$ be a compact operator meaning that the image of a bounded set is a pre-compact set. Then $A(X)$ is separable.*

Proof: If for every $n\in\mathbb{N}$ there is a $1/n$ net for $A(B(0,1))$, then $A(B(0,1))$ would be separable, and a countable dense subset would be the union of these $1/n$ nets. It would follow then that $A(X)$ is also separable. A countable dense subset would be positive rational numbers times these countably many points in $A(B(0,1))$. Therefore, if $A(X)$ is not separable, there is some $\varepsilon>0$ such that there is no ε net, for $A(B(0,1))$ meaning that there are infinitely many points which are all ε apart in $A(B(0,1))$. Let these points be $\{Au_k\}_{k=1}^\infty$. But now this is a contradiction because there can be no convergent subsequence. ■

The big result is called the Hilbert Schmidt theorem. It is a generalization to arbitrary Hilbert spaces of standard finite dimensional results having to do with diagonalizing a symmetric matrix.

Theorem 22.5.3 *Let A be a compact self adjoint operator defined on a Hilbert space H. Then there exists a countable set of eigenvalues, $\{\lambda_i\}$ and an orthonormal set of eigenvectors, u_i satisfying*

$$\lambda_i \text{ is real, } |\lambda_n|\geq|\lambda_{n+1}|,\ Au_i=\lambda_i u_i, \tag{22.15}$$

and either $\lim_{n\to\infty}\lambda_n=0$, or for some n, $\operatorname{span}(u_1,\cdots,u_n)=A(H)$. In any case,

$$\operatorname{span}(\{u_i\}_{i=1}^\infty) \text{ is dense in } A(H). \tag{22.16}$$

and for all $x\in H$, $Ax=\sum_{k=1}^\infty \lambda_k(x,u_k)u_k$ where the sum might be finite.

Proof: First note that if you have a self adjoint operator A, then its eigenvalues are real. Here is why:

$$(Au,u) = \lambda \|u\|^2 = (u,Au) = \bar{\lambda} \|u\|^2.$$

If $\|A\| = 0$ then pick $u \in H$ with $\|u\| = 1$ and let $\lambda = 0$. Since $A(H) = 0$ it follows the span of u is dense in $A(H)$ and the formula for Ax holds.

Assume $A \neq 0$. I will show there exists an eigenvector u. From the definition of $\|A\|$ there exists $x_n, \|x_n\| = 1$, and $\|Ax_n\| \to \|A\| \equiv |\lambda|$. Now it is clear that A^2 is also a compact self adjoint operator. Consider

$$\left(\left(\lambda^2 - A^2\right)x_n, x_n\right) = \lambda^2 (x_n,x_n) - \left(A^2 x_n, x_n\right) = \lambda^2 - \|Ax_n\|^2 \to 0.$$

Since A is compact, there exists a subsequence of $\{x_n\}$ still denoted by $\{x_n\}$ such that Ax_n converges to some element of H. Thus since $\lambda^2 - A^2$ satisfies $\left(\left(\lambda^2 - A^2\right)y, y\right) \geq 0$ in addition to being self adjoint, it follows

$$x, y \to \left(\left(\lambda^2 - A^2\right)x, y\right)$$

satisfies all the axioms for an inner product except for the one which says that $(z,z) = 0$ only if $z = 0$. Therefore, the Cauchy Schwarz inequality may be used to write

$$\left|\left(\left(\lambda^2 - A^2\right)x_n, y\right)\right| \leq \left(\left(\lambda^2 - A^2\right)y, y\right)^{1/2} \left(\left(\lambda^2 - A^2\right)x_n, x_n\right)^{1/2} \leq e_n \|y\|.$$

where $e_n \to 0$ as $n \to \infty$. Taking the sup over $\|y\| \leq 1$, $\lim_{n\to\infty} \left\|\left(\lambda^2 - A^2\right)x_n\right\| = 0$. Since $A^2 x_n$ converges, it follows, since $\lambda \neq 0$ that $\{x_n\}$ is a Cauchy sequence converging to x with $\|x\| = 1$. Therefore, $A^2 x_n \to A^2 x$ and so $\left\|\left(\lambda^2 - A^2\right)x\right\| = 0$. Now this shows that

$$(\lambda I + A)(\lambda I - A)x = 0.$$

If $(\lambda I - A)x = 0$, let $u \equiv x$. If $(\lambda I - A)x = y \neq 0$, let $u \equiv \frac{y}{\|y\|}$. Note that this did not identify the sign of λ. Also note that since $\lambda \neq 0, u \in A(H)$.

Let $\mathscr{A} \in \mathscr{F}$ mean that $\mathscr{A}$ consists of vectors of $A(H), \mathscr{A}$ is an orthonormal set of vectors, and for each $u \in \mathscr{A}, Au = \lambda u$ for some λ. I claim that $\mathscr{A}$ is countable because from the compactness of $A, A(H)$ is separable by Proposition 22.5.2 but these vectors of $A(H)$ are all at least 1/2 apart. Partially order $\mathscr{F}$ by set inclusion. Let $\mathscr{C}$ be a maximal chain. Then $\mathscr{A}_\infty \equiv \cup\mathscr{C}$ is a maximal element of $\mathscr{F}$. I need to show its span is dense in $A(H)$. If $\overline{\text{span}\,\mathscr{A}_\infty}$ fails to contain $A(H)$, then there is a nonzero vector $w \equiv Av$ which is not in $\overline{\text{span}\,\mathscr{A}_\infty}$. Then $(w - Pw)/\|w - Pw\|$ is a unit vector perpendicular to $\overline{\text{span}\,\mathscr{A}_\infty}$. Is this vector in $A(H)$? Is $P(Av) = A(Pv)$? Using Corollary 22.4.4,

$$\begin{aligned} P(w) &= P(Av) = \sum_{k=1}^{N} (Av, u_k) u_k = \sum_{k=1}^{N} \lambda_k (v, u_k) u_k \\ &= \sum_{k=1}^{N} (v, u_k) Au_k = A(Pv) \in A(H) \end{aligned}$$

so this unit vector w is in $\left(\overline{\text{span}\,\mathscr{A}_\infty}\right)^\perp$ and $\mathscr{A}_\infty \cup \{w\}$ is larger than the maximal element of $\mathscr{F}$ so it must be the case that $\overline{\text{span}\,\mathscr{A}_\infty} \supseteq A(H)$ after all.

As noted, this orthonormal set $\mathscr{A}_\infty$ is countable. Let it be $\{u_k\}_{k=1}^N$ where $N \leq \infty$. Thus for $x \in H, Ax \in A(H) \subseteq \overline{A(H)}$ and so, by Corollary 22.4.4,

$$Ax = \sum_{k=1}^{N} (Ax, u_k)\, u_k = \sum_{k=1}^{N} (x, Au_k)\, u_k = \sum_{k=1}^{N} \lambda_k (x, u_k)\, u_k \tag{22.17}$$

and the series converges. Also, the formula implies directly that $Au_m = \lambda_m u_m$ so $|\lambda_m| \leq \|A\|$.

I claim $\limsup_{n\to\infty} |\lambda_n| = 0$. If this were not so, then for some $\varepsilon > 0$, $0 < \varepsilon \leq |\lambda_n|$ for a subsequence still denoted as λ_n but then

$$\|Au_n - Au_m\|^2 = \|\lambda_n u_n - \lambda_m u_m\|^2 = |\lambda_n|^2 + |\lambda_m|^2 \geq 2\varepsilon^2$$

and so there could not exist a convergent subsequence of $\{Au_k\}_{k=1}^\infty$ contrary to the assumption that A is compact. This verifies the claim that $\lim_{n\to\infty} \lambda_n = 0$. Also, since $|\lambda_m| \leq \|A\|$, if $S \subseteq \mathbb{N}$, $\sup_{m\in S} |\lambda_m| = \max_{m\in S} |\lambda_m|$ and so, we can re-number the u_k if necessary such that the eigenvalues satisfy $|\lambda_k| \geq |\lambda_{k+1}|$ for all k. Thus, if $\lambda_m = 0$ for some m, it follows from 22.17 that $A(H)$ is contained in the span of finitely many of the vectors $\{u_k\}$. ■

Define $v \otimes u \in \mathscr{L}(H, H)$ by $v \otimes u(x) = (x, u)\, v$, then 22.17 is of the form

$$A = \sum_{k=1}^{N} \lambda_k u_k \otimes u_k$$

This is the content of the following corollary.

Corollary 22.5.4 *The main conclusion of the above theorem can be written as* $A = \sum_{k=1}^{N} \lambda_k u_k \otimes u_k$ *where the convergence of the partial sums takes place in the operator norm.*

Proof: Without loss of generality, assume $N = \infty$.

$$\begin{aligned} & \left|\left(\left(A - \sum_{k=1}^{n} \lambda_k u_k \otimes u_k\right) x, y\right)\right| = \left|\left(Ax - \sum_{k=1}^{n} \lambda_k (x, u_k)\, u_k, y\right)\right| \\ = & \left|\left(\sum_{k=n}^{\infty} \lambda_k (x, u_k)\, u_k, y\right)\right| = \left|\sum_{k=n}^{\infty} \lambda_k (x, u_k)\, (u_k, y)\right| \\ \leq & |\lambda_n| \left(\sum_{k=n}^{\infty} |(x, u_k)|^2\right)^{1/2} \left(\sum_{k=n}^{\infty} |(y, u_k)|^2\right)^{1/2} \leq |\lambda_n|\, \|x\|\, \|y\| \end{aligned}$$

It follows $\|(A - \sum_{k=1}^{n} \lambda_k u_k \otimes u_k)(x)\| \leq |\lambda_n|\, \|x\|$ ■

Corollary 22.5.5 *Let A be a compact self adjoint operator and*

$$A = \sum_{k=1}^{N} \lambda_k u_k \otimes u_k$$

where $Au_k = \lambda_k u_k$ with $\|u_k\| = 1$, the $|\lambda_k|$ decreasing. Then $|\lambda_{k+1}| = \|A_k\|$ where A_k is the restriction of A to $\{u_1, \ldots, u_k\}^\perp$.

Proof: First note that if $V_k \equiv \{u_1, ..., u_k\}^\perp$, then $A : V_k \to V_k$. Thus

$$\|A_k\| \equiv \sup\{\|A_k u\|, \|u\| \leq 1, (u, u_i) = 0, i \leq k\}$$

Since the $|\lambda_k|$ are decreasing, this will be maximized by picking $u = u_{k+1}$ and then the result is just $|\lambda_k|$. ■

Lemma 22.5.6 *If V_λ is the eigenspace for $\lambda \neq 0$ and $B : V_\lambda \to V_\lambda$ is a compact self adjoint operator with $Bx = \lambda x$ for all $x \in V_\lambda$ then V_λ must be finite dimensional.*

Proof: First note that $B(V_\lambda) = V_\lambda$ because if $u \in V_\lambda$, then $Bu = \lambda u$ and so $B\left(\frac{u}{\lambda}\right) = u$. From Theorem 22.5.3, $\mathscr{A}_\infty$ that maximal set might be finite in which case it would yield a finite orthonormal basis for $V_\lambda = B(V_\lambda)$. But it can't be infinite because there is only one eigenvalue and it is not zero so cannot converge to 0. ■

Next is the case of most interest when H is separable. In this case, the eigenfunctions actually give an orthonormal basis for H.

Corollary 22.5.7 *Let A be a compact self adjoint operator defined on a separable Hilbert space, H. Then there exists a countable set of eigenvalues, $\{\lambda_i\}$ and an orthonormal set of eigenvectors $\{v_i\}$ satisfying $Av_i = \lambda_i v_i, \|v_i\| = 1, \operatorname{span}(\{v_i\}_{i=1}^\infty)$ is dense in H.Furthermore, if $\lambda_i \neq 0$, the space, $V_{\lambda_i} \equiv \{x \in H : Ax = \lambda_i x\}$ is finite dimensional.*

Proof: Let B be the restriction of A to V_{λ_i}. Thus B is a compact self adjoint operator which maps V_λ to V_λ and has only one eigenvalue λ_i on V_{λ_i}. By Lemma 22.5.6, V_λ is finite dimensional. As to the density of some $\operatorname{span}(\{v_i\}_{i=1}^\infty)$ in H, let $W \equiv \overline{\operatorname{span}(\{u_i\})}^\perp$ where $A = \sum_{k=1}^N \lambda_k u_k \otimes u_k$. By Theorem 22.4.2, there is a maximal orthonormal set of vectors, $\{w_i\}_{i=1}^\infty$ whose span is dense in W. There are only countably many of these since the space H is separable. Then consider $\{v_i\}_{i=1}^\infty = \{u_i\}_{i=1}^\infty \cup \{w_i\}_{i=1}^\infty$. $Aw_i = \sum_{k=1}^N \lambda_k (w_i, u_k) = 0$. Thus each w_i is an eigenvector for A. ■

Suppose $\lambda \notin \{\lambda_k\}_{k=1}^\infty$, the eigenvalues of A, and $\lambda \neq 0$. Then the above formula for A, yields an interesting formula for $(A - \lambda I)^{-1}$. Note first that since $\lim_{n\to\infty} \lambda_n = 0$, it follows that $\lambda_n^2/(\lambda_n - \lambda)^2$ must be bounded, say by a positive constant, M.

Corollary 22.5.8 *Let A be a compact self adjoint operator and let $\lambda \notin \{\lambda_n\}_{n=1}^\infty$ and $\lambda \neq 0$ where the λ_n are the eigenvalues of A. $(Ax = \lambda x, x \neq 0)$Then*

$$(A - \lambda I)^{-1} x = -\frac{1}{\lambda} x + \frac{1}{\lambda} \sum_{k=1}^\infty \frac{\lambda_k}{\lambda_k - \lambda} (x, u_k) u_k. \tag{22.18}$$

Proof: Let $m < n$. Then since the $\{u_k\}$ form an orthonormal set,

$$\left| \sum_{k=m}^n \frac{\lambda_k}{\lambda_k - \lambda} (x, u_k) u_k \right| \leq \left(\sum_{k=m}^n \left(\frac{\lambda_k}{\lambda_k - \lambda} \right)^2 |(x, u_k)|^2 \right)^{1/2} \leq M \left(\sum_{k=m}^n |(x, u_k)|^2 \right)^{1/2}. \tag{22.19}$$

But from Bessel's inequality, $\sum_{k=1}^\infty |(x, u_k)|^2 \leq \|x\|^2$ and so for m large enough, the first term in 22.19 is smaller than ε. This shows the infinite series in 22.18 converges. It is now routine to verify that the formula in 22.18 is the inverse. ■

22.5.1 Nuclear Operators

A very useful idea in linear algebra is the trace of a matrix. It is one of the principle invariants and is just the sum of the entries along the main diagonal. Thus if $A \in \mathscr{L}(\mathbb{C}^n, \mathbb{C}^n)$, the trace is $\sum_i \boldsymbol{e}_i^T A \boldsymbol{e}_i$ where the $\boldsymbol{e}_i$ are the standard orthonormal basis vectors, $\boldsymbol{e}_i$ having a 1 in the i^{th} position and 0 elsewhere. If you used another orthonormal basis, $\{\boldsymbol{v}_i\}$ the trace would be the same because the mapping $\boldsymbol{v}_i \to \boldsymbol{e}_i$ will preserve lengths and is therefore a unitary transformation. The two computations would involve a similarity transformation. In infinite dimensions when you have a separable Hilbert space, the notion of trace might not make sense. The nuclear operators are those for which it will make sense.

Definition 22.5.9 *A self adjoint operator $A \in \mathscr{L}(H,H)$ for H a separable Hilbert space is called a nuclear operator if for some complete orthonormal set, $\{e_k\}$,It follows that $\sum_{k=1}^{\infty} |(Ae_k, e_k)| < \infty$.*

We specialize to self adjoint operators because this will ensure that (Ax,x) is real. To begin with here is an interesting lemma.

Lemma 22.5.10 *Suppose $\{A_n\}$ is a sequence of compact operators in $\mathscr{L}(X,Y)$ for two Banach spaces, X and Y and suppose $A \in \mathscr{L}(X,Y)$ and $\lim_{n\to\infty} \|A - A_n\| = 0$. Then A is also compact.*

Proof: Let D be a bounded set in X such that $\|b\| \leq C$ for all $b \in D$. I need to verify $A(B)$ is totally bounded. Suppose then it is not. Then there exists $\varepsilon > 0$ and an infinite sequence, $\{Ab_i\}$ where $b_i \in D$ and $\|Ab_i - Ab_j\| \geq \varepsilon$ whenever $i \neq j$. Then let n be large enough that $\|A - A_n\| \leq \frac{\varepsilon}{4C}$.Then

$$\begin{aligned} \|A_n b_i - A_n b_j\| &= \|Ab_i - Ab_j + (A_n - A) b_i - (A_n - A) b_j\| \\ &\geq \|Ab_i - Ab_j\| - \left(\|(A_n - A) b_i\| + \|(A_n - A) b_j\|\right) \\ &\geq \|Ab_i - Ab_j\| - 2\frac{\varepsilon}{4C} C \geq \frac{\varepsilon}{2}, \end{aligned}$$

a contradiction to A_n being compact. ■

Then one can prove the following lemma. In this lemma, $A \geq 0$ will mean $(Ax,x) \geq 0$.

Lemma 22.5.11 *Let $A \geq 0$ be a nuclear operator defined on a separable Hilbert space H. Then A is compact and also, whenever $\{e_k\}$ is a complete orthonormal set,*

$$A = \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} (Ae_i, e_j)\, e_i \otimes e_j.$$

Proof: First consider the formula. Since A is given to be continuous,

$$Ax = A\left(\sum_{j=1}^{\infty} (x, e_j)\, e_j\right) = \sum_{j=1}^{\infty} (x, e_j) Ae_j,$$

the series converging because $x = \sum_{j=1}^{\infty} (x, e_j)\, e_j$. Then also since A is self adjoint,

$$\sum_{j=1}^{\infty} \sum_{i=1}^{\infty} (Ae_i, e_j)\, e_i \otimes e_j\,(x) \equiv \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} (Ae_i, e_j)\,(x, e_j)\, e_i = \sum_{j=1}^{\infty} (x, e_j) \sum_{i=1}^{\infty} (Ae_i, e_j)\, e_i$$

$$= \sum_{j=1}^{\infty} (x,e_j) \sum_{i=1}^{\infty} (Ae_j,e_i)\, e_i = \sum_{j=1}^{\infty} (x,e_j)\, Ae_j$$

Next consider the claim that A is compact. Let $C_A \equiv \left(\sum_{j=1}^{\infty} \left|(Ae_j,e_j)\right|\right)^{1/2}$. Let A_n be defined by $A_n \equiv \sum_{j=1}^{\infty}\sum_{i=1}^{n} (Ae_i,e_j)\,(e_i \otimes e_j)$. Then A_n has values in $\operatorname{span}(e_1,\cdots,e_n)$ and so it must be a compact operator because bounded sets in a finite dimensional space must be precompact. Then

$$|(Ax - A_n x, y)| = \left| \sum_{j=1}^{\infty} \sum_{i=n+1}^{\infty} (Ae_i,e_j)\,(y,e_j)\,(e_i,x) \right| = \left| \sum_{j=1}^{\infty} (y,e_j) \sum_{i=n+1}^{\infty} (Ae_i,e_j)\,(e_i,x) \right|$$

$$\leq \left| \sum_{j=1}^{\infty} \left|(y,e_j)\right| (Ae_j,e_j)^{1/2} \sum_{i=n+1}^{\infty} (Ae_i,e_i)^{1/2} |(e_i,x)| \right|$$

$$\leq \left(\sum_{j=1}^{\infty} \left|(y,e_j)\right|^2 \right)^{1/2} \left(\sum_{j=1}^{\infty} \left|(Ae_j,e_j)\right| \right)^{1/2} \cdot \left(\sum_{i=n+1}^{\infty} |(x,e_i)|^2 \right)^{1/2} \left(\sum_{i=n+1}^{\infty} |(Ae_ie_i)| \right)^{1/2}$$

$$\leq \|y\|\,\|x\|\, C_A \left(\sum_{i=n+1}^{\infty} |(Ae_i,e_i)| \right)^{1/2}$$

and this shows that if n is sufficiently large, it follows that $|((A-A_n)x,y)| \leq \varepsilon \|x\|\,\|y\|$ so for such n, $\|A-A_n\| < \varepsilon$. Therefore, $\lim_{n\to\infty} \|A-A_n\| = 0$ and so A is the limit in operator norm of finite rank bounded linear operators, each of which is compact. Therefore, A is also compact. ■

Definition 22.5.12 *The trace of a nuclear operator $A \in \mathscr{L}(H,H)$ such that $A \geq 0$ is defined to equal $\sum_{k=1}^{\infty} (Ae_k,e_k)$ where $\{e_k\}$ is an orthonormal basis for the separable Hilbert space, H.*

Theorem 22.5.13 *Definition 22.5.12 is well defined and equals $\sum_{j=1}^{\infty} \lambda_j$ where the λ_j are the nonnegative eigenvalues of A.*

Proof: Suppose $\{u_k\}$ be the basis of eigenvectors of the Hilbert Schmidt theorem Then $e_k = \sum_{j=1}^{\infty} u_j\,(e_k,u_j)$. By Lemma 22.5.11 A is compact and so by the Hilbert Schmidt theorem, Theorem 22.5.3, $A = \sum_{k=1}^{\infty} \lambda_k u_k \otimes u_k$ where the u_k are the orthonormal eigenvectors of A which form a complete orthonormal set. Then

$$\sum_{k=1}^{\infty} (Ae_k,e_k) = \sum_{k=1}^{\infty} \left(A(e_k), \sum_{i=1}^{\infty} u_i\,(e_k,u_i) \right) = \sum_{k=1}^{\infty}\sum_{i=1}^{\infty} (A(e_k),u_i)\,(e_k,u_i)$$

$$= \sum_{k=1}^{\infty}\sum_{i=1}^{\infty} (e_k,Au_i)\,(e_k,u_i) = \sum_{k=1}^{\infty}\sum_{i=1}^{\infty} \left(e_k, \sum_{j=1}^{\infty} (Au_i,u_j)\,u_j \right) (e_k,u_i)$$

$$= \sum_{k=1}^{\infty}\sum_{i}\sum_{j} (Au_i,u_j)\,(e_k,u_j)\,(u_i,e_k) = \sum_{k=1}^{\infty}\sum_{i}\sum_{j} \lambda_j \delta_{ij}\,(e_k,u_j)\,(u_i,e_k)$$

$$= \sum_{k=1}^{\infty}\sum_{i=1}^{\infty} \lambda_i |(u_i,e_k)|^2 = \sum_{i=1}^{\infty} \lambda_i \sum_{k=1}^{\infty} |(u_i,e_k)|^2 = \sum_{i=1}^{\infty} \lambda_i \|u_i\|^2 = \sum_{i=1}^{\infty} \lambda_i \text{ ■}$$

This is just like it is for a matrix. Recall the trace of a matrix is the sum of the eigenvalues.

It is also easy to see that in any Hilbert space, there exist nuclear operators. Let $\sum_{k=1}^{\infty}|\lambda_k|<\infty$. Then let $\{e_k\}$ be an orthonormal set of vectors. Let $A\equiv\sum_{k=1}^{\infty}\lambda_k e_k\otimes e_k, \lambda_k$ real. It is not too hard to verify this works.

Much more can be said about nuclear operators.

22.5.2 Hilbert Schmidt Operators

In all of this, H, G will be separable Hilbert spaces.

Definition 22.5.14 *Let T be a continuous linear mapping from H to G and whenever $\{e_k\}$ is an orthonormal basis for H, then $\sum_k \|Te_k\|^2<\infty$. Such an operator is called a Hilbert Schmidt operator.We write $T\in\mathscr{L}_2(H,G)$. Picking an orthonormal basis (complete orthonormal set), define $\|T\|_{\mathscr{L}_2}^2\equiv\sum_k\|Te_k\|_G^2$.*

It is necessary to show that this is well defined and does not depend on the orthonormal basis. For now let the orthonormal basis be fixed.

Lemma 22.5.15 *If T is Hilbert Schmidt, then $\|T\|_{\mathscr{L}(H,G)}\leq\|T\|_{\mathscr{L}_2(H,G)}$. Also T^* is Hilbert Schmidt and T^*T is Hilbert Schmidt. In fact, if $T\in\mathscr{L}_2(H,G)$ and $S\in\mathscr{L}(G,H)$ then $ST\in\mathscr{L}_2(H,H)$.*

Proof: Pick an orthonormal basis for $H,\{e_k\}$ $\sum_k\|Te_k\|^2<\infty$ and an orthonormal basis for $G,\{f_k\}$. Then let $x=\sum_{k=1}^n(x,e_k)e_k\equiv\sum_{k=1}^n x_k e_k$. Then

$$\|Tx\|=\left(\sum_{k=1}^{\infty}|(Tx,f_k)|^2\right)^{1/2}=\left(\sum_{k=1}^{\infty}\left|\left(\sum_{j=1}^{n}x_jTe_j,f_k\right)\right|^2\right)^{1/2}$$

$$\begin{aligned}&=\left(\sum_{k=1}^{\infty}\left|\sum_{j=1}^{n}(x_jTe_j,f_k)\right|^2\right)^{1/2}\leq\sum_{j=1}^{n}\left(\sum_{k=1}^{\infty}\left|(x_jTe_j,f_k)\right|^2\right)^{1/2}\\&\leq\sum_{j=1}^{n}|x_j|\left(\sum_{k=1}^{\infty}\left|(Te_j,f_k)\right|^2\right)^{1/2}\end{aligned}$$

$$=\sum_{j=1}^{n}|x_j|\left\|Te_j\right\|\leq\left(\sum_{j=1}^{n}|x_j|^2\right)^{1/2}\left(\sum_{j=1}^{n}\left\|Te_j\right\|^2\right)^{1/2}=\|x\|\,\|T\|_{\mathscr{L}_2}$$

Therefore, since finite sums of the form $\sum_{k=1}^n x_k e_k$ are dense in H, it follows $\|T\|\leq\|T\|_{\mathscr{L}_2}$.

Letting $\{f_i\}$ be orthonormal in G, $\|Te_k\|^2=\sum_j\left|(Te_k,f_j)\right|^2$ and so

$$\sum_k\|Te_k\|^2=\sum_k\sum_j\left|(Te_k,f_j)\right|^2=\sum_j\sum_k\left|(e_k,T^*f_j)\right|^2=\sum_j\left\|T^*f_j\right\|^2$$

so T^* is also Hilbert Schmidt.

Let $\{e_i\}$ be an orthonormal basis for H so $\sum_i \|Te_i\|^2 < \infty$. Then

$$\sum_i \|T^*Te_i\|^2 \leq \|T^*\|^2 \sum_i \|Te_i\|^2 < \infty$$

For the last claim, let $\{e_k\}$ be an orthonormal basis. Then

$$\sum_k \|ST(e_k)\|^2 \leq \sum_k \|S\|^2 \|Te_k\|^2 < \infty. \blacksquare$$

Definition 22.5.16 *Define $(S,T) \equiv \sum_k (Se_k, Te_k)$ where $\{e_k\}$ is a given orthonormal basis. This is well defined because the sum converges absolutely. Indeed,*

$$\sum_k |(Se_k, Te_k)| \leq \sum_k |Se_k| |Te_k| \leq \left(\sum_k |Se_k|^2\right)^{1/2} \left(\sum_k |Te_k|^2\right)^{1/2} < \infty$$

Definition 22.5.17 *For $X \in G$ and $Y \in H, X \otimes Y(h) \equiv X(h,Y)$. This is a continuous linear map from H to G because $\|X \otimes Y(h)\| = \|X(h,Y)\| \leq \|h\| \|Y\| \|X\|$. Next we show $X \otimes Y \in \mathscr{L}_2(H,G)$ among other things.*

Theorem 22.5.18 *$\mathscr{L}_2(H,G)$ is a separable Hilbert space with norm given by*

$$\|T\|_{\mathscr{L}_2} \equiv \left(\sum_k \|Te_k\|^2\right)^{1/2}$$

where $\{e_k\}$ is some fixed orthonormal basis for H. Also $\mathscr{L}_2(H,G) \subseteq \mathscr{L}(H,G)$ and

$$\|T\| \leq \|T\|_{\mathscr{L}_2}. \tag{22.20}$$

All Hilbert Schmidt opearators are compact. Also for $X \in G$ and $Y \in H, X \otimes Y \in \mathscr{L}_2(H,G)$ and

$$\|X \otimes Y\|_{\mathscr{L}_2} = \|X\|_G \|Y\|_H \tag{22.21}$$

If T is Hilbert Schmidt, then so is T^*T and $T^*$and ST for any $S \in \mathscr{L}(G,H)$. If $T = T^*$ and $G = H$, then the choice of orthonormal basis in computing $\|T\|_{\mathscr{L}_2}$ is not important.

Proof: Is $\|T\|_{\mathscr{L}_2}$ really a norm? This obviously is so except for the triangle inequality. But this follows from the triangle inequality.

$$\begin{aligned} \|T+S\|_{\mathscr{L}_2} &\equiv \left(\sum_k \|Te_k + Se_k\|^2\right)^{1/2} \leq \left(\sum_k \|Te_k\|^2\right)^{1/2} + \left(\sum_k \|Se_k\|^2\right)^{1/2} \\ &= \|T\|_{\mathscr{L}_2} + \|S\|_{\mathscr{L}_2} \end{aligned}$$

Next is the claim that $\mathscr{L}_2$ is a Hilbert space. So pick an orthonormal basis $\{e_k\}$. It is clear that $\mathscr{L}_2$ is an inner product space with respect to the inner product described above in Definition 22.5.16.

Consider completeness. Suppose that $\{T_n\}$ is a Cauchy sequence in $\mathscr{L}_2(H,G)$. Then from 22.20 $\{T_n\}$ is a Cauchy sequence in $\mathscr{L}(H,G)$ and so there exists a unique T such

that $\lim_{n\to\infty}\|T_n - T\| = 0$. Then it only remains to verify $T \in \mathscr{L}_2(H,G)$. But by Fatou's lemma, for $\{e_k\}$ orthonormal,

$$\sum_k \|Te_k\|^2 \le \lim\inf_{n\to\infty}\sum_k \|T_n e_k\|^2 \equiv \lim\inf_{n\to\infty}\|T_n\|^2_{\mathscr{L}_2} < \infty.$$

All that remains is to verify $\mathscr{L}_2(H,G)$ is separable and these Hilbert Schmidt operators are compact. I will show an orthonormal basis for $\mathscr{L}_2(H,G)$ is $\{f_j \otimes e_k\}$ where $\{f_k\}$ is an orthonormal basis for G and $\{e_k\}$ is an orthonormal basis for H. Here, for $f \in G$ and $e \in H, f\otimes e(x) \equiv (x,e) f$.

I need to show $f_j \otimes e_k \in \mathscr{L}_2(H,G)$ and that it is an orthonormal basis for $\mathscr{L}_2(H,G)$ as claimed. Let the $\{e_k\}$ be the orthonormal basis used to define the inner product but the $\{f_j\}$ are just an arbitrary orthonormal basis for G.

$$\sum_k \|f_j \otimes e_i(e_k)\|^2 = \sum_k \|f_j \delta_{ik}\|^2 = \|f_j\|^2 = 1 < \infty$$

so each of these operators is in $\mathscr{L}_2(H,G)$. As noted above, they are also each continuous. Next I show they are orthonormal. From the definition of the inner product,

$$\begin{aligned}(f_j\otimes e_k, f_s \otimes e_r) &= \sum_p (f_j \otimes e_k(e_p), f_s \otimes e_r(e_p)) \\ &= \sum_p \delta_{rp}\delta_{kp}(f_j,f_s) = \sum_p \delta_{rp}\delta_{kp}\delta_{js}\end{aligned}$$

If $j = s$ and $k = r$ this reduces to 1. Otherwise, this gives 0. Thus these operators are orthonormal.

Why is $\mathscr{L}_2(H,G)$ a separable Hilbert space? Let $T \in \mathscr{L}_2(H,G)$. Consider

$$T_n \equiv \sum_{i=1}^n \sum_{j=1}^n (Te_i, f_j) f_j \otimes e_i$$

Then

$$T_n e_k = \sum_{i=1}^n \sum_{j=1}^n (Te_i, f_j)(e_k, e_i) f_j = \sum_{j=1}^n (Te_k, f_j) f_j,$$

a partial sum for Te_k. It follows $\|T_n e_k\| \le \|Te_k\|$ and $\lim_{n\to\infty} T_n e_k = Te_k$. Therefore, from the dominated convergence theorem,

$$\lim_{n\to\infty}\|T - T_n\|^2_{\mathscr{L}_2} \equiv \lim_{n\to\infty}\sum_k \|(T-T_n)e_k\|^2 = 0.$$

Therefore, the linear combinations of the $f_j \otimes e_i$ are dense in $\mathscr{L}_2(H,G)$ and this proves completeness of the orthonomal basis.

By only using rational scalars in the linear combinations we see that $\mathscr{L}_2(H,G)$ is separable. From 22.20 it also shows that every $T \in \mathscr{L}_2(H,G)$ is the limit in the operator norm of a sequence of compact operators. This follows because each of the $f_j \otimes e_i$ is easily seen to be a compact operator because if $B \subseteq H$ is bounded, then $(f_j \otimes e_i)(B)$ is a bounded subset of a one dimensional vector space so it is pre-compact. Thus T_n is compact, being a finite sum of these. By Lemma 22.5.10, so is T.

Consider 22.21.

$$\|X\otimes Y\|^2_{\mathscr{L}_2} \equiv \sum_k \|X\otimes Y(e_k)\|^2_G \equiv \sum_k \|X(e_k,Y)\|^2_G$$

$$= \|X\|^2_G \sum_k |(e_k,Y)|^2 = \|X\|^2_G \|Y\|^2_H$$

Finally, consider the last claim. I need to show that for self adjoint operators in $\mathscr{L}_2$ the choice of orthonormal basis does not matter. This is because if $\{e_k\}, \{f_j\}$ are two orthonormal bases, then

$$\sum_k \|Te_k\|^2 = \sum_k \sum_j \left|(Te_k, f_j)\right|^2 = \sum_j \sum_k \left|(e_k, Tf_j)\right|^2 = \sum_j \left\|Tf_j\right\|^2. \blacksquare$$

In fact the orthonormal basis does not matter in defining the norm of any Hilbert Schmidt operator which is not surprising from linear algebra. I will show this as an application a little later in Proposition 22.6.4.

22.6 Roots of Positive Linear Maps

In this section, H will be a Hilbert space, real or complex, and T will denote an operator which satisfies the following definition. This will be a more general result than the above because it will hold for infinite dimensional spaces.

Definition 22.6.1 *Let T satisfy $T = T^*$ (Hermitian) and for all $x \in H$,*

$$(Tx,x) \geq 0 \tag{22.22}$$

Such an operator is referred to as positive and self adjoint. It is probably better to refer to such an operator as "nonnegative" since the possibility that $Tx = 0$ for some $x \neq 0$ is not being excluded. Instead of "self adjoint" you can also use the term, Hermitian. To save on notation, write $T \geq 0$ to mean T is positive, satisfying 22.22. When we say $A \leq B$ this means $B - A \geq 0$.

A useful theorem about the existence of roots of positive self adjoint operators is presented. This proof is very elementary. I found it in [34] for square roots.

22.6.1 The Product of Positive Self Adjoint Operators

With the above definition here is a fundamental result about positive self adjoint operators.

Proposition 22.6.2 *Let S, T be positive and self adjoint such that $ST = TS$. Then ST is also positive and self adjoint.*

Proof: It is obvious that ST is self adjoint.

$$(STx,y) = (TSx,y) = (Sx,Ty) = (x,STy)$$

The only problem is to show that ST is positive. The idea is to write $S = S_{n+1} + \sum_{k=0}^n S_k^2$ where $S_0 = S$ and the operators S_k are self adjoint. This is because if you have $\left(TS^2x,x\right)$,

where everything commutes, this equals $(STSx,x) = (TSx,Sx) \geq 0$. Thus it will be possible to deal with the terms of the sum which are squared. First assume $(Sx,x) \leq (x,x)$ so $S \leq I$.

Define a sequence recursively as follows.

$$S_{n+1} = S_n - S_n^2,\ S \equiv S_0 \tag{22.23}$$

Then $\sum_{k=0}^{n} S_k^2 = \sum_{k=0}^{n} (S_k - S_{k+1}) = S - S_{n+1},\ S = S_{n+1} + \sum_{k=0}^{n} S_k^2$. Now $S_0 \geq 0$ by assumption. Assume $S_n \geq 0$. Then

$$S_{n+1} = S_n - S_n^2 = (I - S_n) S_n (S_n + (I - S_n)) = S_n^2 (I - S_n) + (I - S_n)^2 S_n$$

It follows that $S_{n+1} \geq 0$ because clearly those two terms on the end are positive. Therefore,

$$(Sx,x) = (S_{n+1}x,x) + \sum_{k=0}^{n} \left(S_k^2 x,x\right) \geq \sum_{k=0}^{n} \|S_k x\|^2,\ (Sx,x) \geq \sum_{k=0}^{\infty} \|S_k x\|^2$$

also and so $\lim_{k\to\infty} \|S_k x\| = 0$. $TSx = TS_{n+1}x + \sum_{k=0}^{n} TS_k^2 x$.

$$(TSx,x) = (S_{n+1}x,Tx) + \sum_{k=0}^{n} \left(TS_k^2 x,x\right) = (S_{n+1}x,Tx) + \sum_{k=0}^{n} (TS_k x, S_k x)$$

so passing to a limit as $n \to \infty, (TSx,x) = 0 + \limsup_{n\to\infty} \sum_{k=0}^{n} (TS_k x, S_k x) \geq 0$.

Thus if $S \leq I$, the theorem is proved. If S is general, $\frac{S}{\|S\|} \leq I$. In this case, $\left(T\frac{S}{\|S\|}x,x\right) = \left(\frac{S}{\|S\|}Tx,x\right) \geq 0$ and so $(STx,x) \geq 0$. ■

The proposition is like the familiar statement about real numbers which says that when you multiply two nonnegative real numbers the result is a nonnegative real number.

22.6.2 Roots of Positive Self Adjoint Operators

With this preparation, it is time to give the theorem about roots.

Theorem 22.6.3 *Let $T \in \mathscr{L}(H,H)$ be a positive self adjoint linear operator. Then for $m \in \mathbb{N}$, there exists a unique m^{th} root A with the following properties. $A^m = T, A$ is positive and self adjoint, A commutes with every operator which commutes with T.*

Proof: Define the following sequence of operators:

$$A_0 \equiv 0,\ A_{n+1} \equiv A_n + \frac{1}{m}(T - A_n^m)$$

Say $T \leq I$.

Claim 1: $A_n \leq I$.

Proof of Claim 1: True if $n = 0$. Assume true for n. Then

$$\begin{aligned} I - A_{n+1} &= I - A_n + \frac{1}{m}(A_n^m - T) \geq I - A_n + \frac{1}{m}(A_n^m - I) \\ &= I - A_n - \frac{1}{m}(I - A_n^m) \\ &= (I - A_m) - \frac{1}{m}(I - A_m)\left(I + \cdots + A_n^{m-1}\right) \end{aligned}$$

Now, since $A_n \leq I, I + \cdots + A_n^{m-1} \leq mI$, it follows that

$$= (I - A_m)\left(I - \frac{1}{m}\left(I + \cdots + A_n^{m-1}\right)\right) \geq (I - A_m)(I - I) = 0$$

so by induction, $A_n \leq I$.

Claim 2: $A_n \leq A_{n+1}$.

Proof of Claim 2: From the definition of A_n, this is true if $n = 0$ because

$$A_1 = T \geq 0 = A_0.$$

Suppose true for n. Then from Claim 1,

$$\begin{aligned} A_{n+2} - A_{n+1} &= A_{n+1} + \frac{1}{m}\left(T - A_{n+1}^m\right) - \left[A_n + \frac{1}{m}\left(T - A_n^m\right)\right] \\ &= A_{n+1} - A_n + \frac{1}{m}\left(A_n^m - A_{n+1}^m\right) \end{aligned}$$

$$= (A_{n+1} - A_n) - (A_{n+1} - A_n)\frac{1}{m}\left(A_{n+1}^{m-1} + A_{n+1}^{m-2}A_n + \cdots + A_n^{m-1}\right)$$

$$\geq (A_{n+1} - A_n) - (A_{n+1} - A_n)I = 0$$

since each $A_n, A_{n+1} \leq I$, so this proves the claim.

Claim 3: $A_n \geq 0$

Proof of Claim 3: This is true if $n = 0$. Suppose it is true for n.

$$\begin{aligned} (A_{n+1}x, x) &= (A_n x, x) + \frac{1}{m}(Tx, x) - \frac{1}{m}(A_n^m x, x) \\ &\geq (A_n x, x) + \frac{1}{m}(Tx, x) - \frac{1}{m}(A_n x, x) \geq 0 \end{aligned}$$

because by Proposition 22.6.2, $A_n - A_n^m = A_n\left(I - A_n^{m-1}\right) \geq 0$ because $A_n \leq I$.

Thus $(A_n x, x)$ is increasing and bounded above so it converges. Now let $n > k$. Using Proposition 22.6.2 $A_n A_k \geq A_k^2$ and also

$$(A_n - A_k)(A_n + A_k) \leq 2(A_n - A_k).$$

Thus the following holds.

$$\|A_n x - A_k x\|^2 = \left(\left(A_n - A_k\right)^2 x, x\right) = \left(A_n^2 x, x\right) - 2\left(A_n A_k x, x\right) + \left(A_k^2 x, x\right)$$

$$\leq \left(A_n^2 x, x\right) - 2\left(A_k^2 x, x\right) + \left(A_k^2 x, x\right) = \left(\left(A_n - A_k\right)\left(A_n + A_k\right)x, x\right)$$

$$\leq 2\left[\left(A_n x, x\right) - \left(A_k x, x\right)\right]$$

which converges to 0 as $k, n \to \infty$. Therefore, $\lim_{n\to\infty} A_n x$ exists since $\{A_n x\}$ is a Cauchy sequence. Let this limit be Ax. Then clearly A is linear. Also, since each $A_n \geq 0$ and self adjoint, the Cauchy Schwarz inequality implies

$$|(Ax, y)| = \lim_{n\to\infty} |(A_n x, y)| \leq \limsup_{n\to\infty} \left|\left(A_n x, x\right)^{1/2}\left(A_n y, y\right)^{1/2}\right| \leq \|x\|\,\|y\|$$

so A is also continuous. Now $(Ax,x)=\lim_{n\to\infty}(A_nx,x)\geq 0$ so A is positive and it is clearly also self adjoint since each A_n is. From passing to the limit in the definition of A_n,

$$Ax=Ax+\frac{1}{m}\left(Tx-A^mx\right)$$

and so $Tx=A^mx$. This proves the theorem in the case that $T\leq I$. Then if $T>I$, consider $T/\|T\|$. $T/\|T\|\leq I$ and so there is B such that $B^m=T/\|T\|$. Let $A=\|T\|^{1/m}B$. This proves the existence of the m^{th} root. It is clear that A commutes with every continuous linear operator that commutes with T because this is true of each of the iterates. In fact, each of these is just a polynomial in T. It remains to verify uniqueness.

Next suppose both A and B are m^{th} roots of T having all the properties stated in the theorem. Then $AB=BA$ because both A and B commute with every operator which commutes with T. Then from Proposition 22.6.2,

$$\left(\left(A^{m-1}+A^{m-2}B+...+B^{m-1}\right)(A-B)x,(A-B)x\right)\geq 0 \tag{22.24}$$

Therefore, $((A^m-B^m)x,(A-B)x)=(0,(A-B)x)=0$.

Now this means $\left(A^kB^l(A-B)x,(A-B)x\right)=0$ for all $k+l=m-1$ since the sum of such terms is 0 and each of them is nonnegative. Now this implies

$$\left(\sqrt{A^kB^l}(A-B)x,\sqrt{A^kB^l}(A-B)x\right)=0$$

and so $\sqrt{A^kB^l}(A-B)x=0\Rightarrow A^kB^l(A-B)x=0,k+l=m-1$. Then, using the binomial theorem,

$$0=\sum_{j=0}^{m-1}\binom{m-1}{j}A^{m-1-j}B^j(-1)^j\overbrace{(A-B)x}^{=(A-B)^{m-1}}=(A-B)^mx$$

This clearly implies $A=B$. To see this, consider $m=7$.

If $m=7,(A-B)^7x=0$ so

$$(A-B)^8x=0$$

so $\left((A-B)^4x,(A-B)^4x\right)=0$ which implies $(A-B)^4x=0$ which implies

$$\left((A-B)^2x,(A-B)^2x\right)=0$$

so $\left((A-B)^2x,x\right)=0$ which yields $((A-B)x,(A-B)x)=0$, so $(A-B)=0$. ■

This next was shown earlier, but this is a nice way to think of it in terms of a square root.

Proposition 22.6.4 *Let $T\in\mathscr{L}_2(H,G)$. Then if $\{e_k\},\{f_j\}$ are two orthonormal bases, then $\sum_k\|Te_k\|^2=\sum_k\|Tf_k\|^2$.*

Proof: T^*T is self adjoint and in $\mathscr{L}_2(H,H)$. Therefore,

$$\sum_k\left\|\sqrt{T^*T}e_k\right\|^2=\sum_k(T^*Te_k,e_k)=\sum_k\|Te_k\|^2$$

which is finite. Thus $\sqrt{T^*T}$ is self adjoint and in $\mathscr{L}_2(H,H)$ and so

$$\sum_k \left\| \sqrt{T^*T} e_k \right\|^2 = \sum_k \left\| \sqrt{T^*T} f_k \right\|^2 = \sum_k \|Tf_k\|^2$$

showing that $\sum_k \|Te_k\|^2 = \sum_k \|Tf_k\|^2$. ■

There is a whole book on powers of operators. This has just given a short introduction. See [33].

22.7 Differential Equations in Banach Space

Here we consider the initial value problem for functions which have values in a Banach space. Let X be a Banach space.

Definition 22.7.1 *Define $BC([a,b];X)$ as bounded continuous functions f which have values in the Banach space X. For $f \in BC([a,b];X)$, γ a real number. Then*

$$\|f\|_\gamma \equiv \sup_{t\in[a,b]} \left\| f(t) e^{\gamma(t-a)} \right\| \tag{22.25}$$

Then this is a norm. The usual norm is given by $\|f\| \equiv \sup_{t\in[a,b]} \|f(t)\|$.

Lemma 22.7.2 *$\|\cdot\|_\gamma$ is a norm for $BC([a,b];X)$ and $BC([a,b];X)$ is a complete normed linear space. Also, a sequence is Cauchy in $\|\cdot\|_\gamma$ if and only if it is Cauchy in $\|\cdot\|$.*

Proof: First consider the claim about $\|\cdot\|_\gamma$ being a norm. To simplify notation, let $T=[a,b]$. It is clear that $\|f\|_\gamma = 0$ if and only if $f=0$ and $\|f\|_\gamma \geq 0$. Also,

$$\|\alpha f\|_\gamma \equiv \sup_{t\in T} \left\| \alpha f(t) e^{\gamma(t-a)} \right\| = |\alpha| \sup_{t\in T} \left\| f(t) e^{\gamma(t-a)} \right\| = |\alpha| \|f\|_\gamma$$

so it does what is should for scalar multiplication. Next consider the triangle inequality.

$$\begin{aligned} \|f+g\|_\gamma &= \sup_{t\in T} \left\| (f(t)+g(t)) e^{\gamma(t-a)} \right\| \leq \sup_{t\in T} \left(\left| f(t) e^{\gamma(t-a)} \right| + \left| g(t) e^{\gamma(t-a)} \right| \right) \\ &\leq \sup_{t\in T} \left| f(t) e^{\gamma(t-a)} \right| + \sup_{t\in T} \left| g(t) e^{\gamma(t-a)} \right| = \|f\|_\gamma + \|g\|_\gamma \end{aligned}$$

The rest follows from the next inequalities.

$$\|f\| \equiv \sup_{t\in T} \|f(t)\| = \sup_{t\in T} \left\| f(t) e^{\gamma(t-a)} e^{-\gamma(t-a)} \right\| \leq e^{|\gamma(b-a)|} \|f\|_\gamma$$

$$\equiv e^{|\gamma(b-a)|} \sup_{t\in T} \left\| f(t) e^{\gamma(t-a)} \right\| \leq \left(e^{|\gamma|(b-a)} \right)^2 \sup_{t\in T} \|f(t)\| = \left(e^{|\gamma|(b-a)} \right)^2 \|f\| \text{ ■}$$

Now consider the ordinary initial value problem

$$x'(t) = F(t, x(t)),\ x(t_0) = x_0,\ t \in [a,b], t_0 \in [a,b] \tag{22.26}$$

where here $F : [a,b] \times X \to X$ is continuous and satisfies the Lipschitz condition

$$\|F(t,x) - F(t,y)\| \leq K \|x-y\|,\ F : [a,b] \times X \to X \text{ is continuous} \tag{22.27}$$

Thanks to the fundamental theorem of calculus, there exists a solution to 22.26 if and only if it is a solution to the integral equation

$$x(t) = x_0 + \int_{t_0}^{t} F(s, x(s))\, ds \tag{22.28}$$

Then we have the following theorem.

Theorem 22.7.3 *Let 22.27 hold. Then there exists a unique solution to 22.26 in* $BC([a,b];X)$.

Proof: Use the norm of 22.25 where $\gamma \neq 0$ is described later. Let $T : BC([a,b];X) \to BC([a,b];X)$ be defined by

$$Tx(t) \equiv x_0 + \int_{t_0}^{t} F(s, x(s))\, ds$$

Then

$$\|Tx(t) - Ty(t)\|_X = \left\| \int_{t_0}^{t} F(s, x(s))\, ds - \int_{t_0}^{t} F(s, y(s))\, ds \right\|$$

$$\leq K \int_{t_0}^{t} \|x(s) - y(s)\|\, ds = K \int_{t_0}^{t} \left\| (x(s) - y(s))\, e^{\gamma(s-a)} e^{-\gamma(s-a)} \right\| ds$$

$$\leq K \int_{t_0}^{t} e^{-\gamma(s-a)} ds \, \|x - y\|_\gamma = K \left(\frac{e^{-\gamma(t-a)}}{-\gamma} + \frac{e^{-\gamma(t_0-a)}}{\gamma} \right) \|x - y\|_\gamma$$

Therefore, letting $\gamma < 0$

$$e^{\gamma(t-a)} \|Tx(t) - Ty(t)\|_X \leq K \left(\frac{1}{-\gamma} + \frac{e^{\gamma(t-t_0)}}{\gamma} \right) \|x - y\|_\gamma < K \left(\frac{1}{|\gamma|} \right) \|x - y\|_\gamma$$

$$\|Tx - Ty\|_\gamma \leq K \left(\frac{1}{|\gamma|} \right) \|x - y\|_\gamma$$

Letting $\gamma = -m^2$, this reduces to

$$\|Tx - Ty\|_{-m^2} \leq \frac{K}{m^2} \|x - y\|_{-m^2}$$

and so if $K/m^2 < 1/2$, this shows the solution to the integral equation is the unique fixed point of a contraction mapping defined on $BC([a,b];X)$. This shows existence and uniqueness of the initial value problem 22.26. ■

Definition 22.7.4 *Let* $S : [0,\infty) \to \mathscr{L}(X,X)$ *be continuous and satisfy*

1. $S(t+s) = S(t)S(s)$ *called the semigroup identity.*
2. $S(0) = I$
3. $\lim_{h \to 0+} \frac{S(h)x - x}{h} = Ax$ *for* A *a densely defined closed linear operator whenever* $x \in D(A) \subseteq X$.

Then S is called a continuous semigroup and A is said to generate S.

Then we have the following corollary of Theorem 22.7.3. First note the following. For $t \geq 0$ and $h \geq 0$, if $x \in D(A)$, the semigroup identity implies

$$\lim_{h\to 0} \frac{S(t+h)x - S(t)x}{h} = \lim_{h\to 0} S(t) \frac{S(h)x - x}{h} = S(t) \lim_{h\to 0} \frac{S(h)x - x}{h} \equiv S(t)Ax$$

As shown above, $\mathscr{L}(X,X)$ is a Banach space with the operator norm whenever X is a Banach space.

Corollary 22.7.5 *Let X be a Banach space and let $A \in \mathscr{L}(X,X)$. Let $S(t)$ be the solution in $\mathscr{L}(X,X)$ to*

$$S'(t) = AS(t),\ S(0) = I, t \geq 0 \tag{22.29}$$

Then $t \to S(t)$ is a continuous semigroup whose generator is A. In this case that A is actually defined on all of X, not just on a dense subset. Furthermore, in this case where $A \in \mathscr{L}(X,X)$, $S(t)A = AS(t)$. If $T(t)$ is any semigroup having A as a generator, then $T(t) = S(t)$. Also you can express $S(t)$ as a power series, $S(t) = \sum_{n=0}^{\infty} \frac{(At)^n}{n!}$.

Proof: The solution to the initial value problem 22.29 exists on $[-b,b]$ for all b so it exists on all of $\mathbb{R}$ thanks to the uniqueness on every finite interval. First consider the semigroup property. Let $\Psi(t) \equiv S(t+s), \Phi(t) \equiv S(t)S(s)$. Then

$$\Psi'(t) = S'(t+s) = AS(t+s) = A\Psi(t),\ \Psi(0) = S(s)$$

$$\Phi'(t) = S'(t)S(s) = AS(t)S(s) = A\Phi(t),\ \Phi(0) = S(s)$$

By uniqueness, $\Phi(t) = \Psi(t)$ for all $t \geq 0$. Thus $S(t)S(s) = S(t+s) = S(s)S(t)$. Now from this, for $t > 0$

$$S(t)A = S(t) \lim_{h\to 0} \frac{S(h) - I}{h} = \lim_{h\to 0} S(t) \frac{S(h) - I}{h} = \lim_{h\to 0} \frac{S(h) - I}{h} S(t) = AS(t).$$

As to A being the generator of $S(t)$, letting $x \in X$, then from the differential equation solved,

$$\lim_{h\to 0+} \frac{S(h)x - x}{h} = \lim_{h\to 0+} \frac{1}{h} \int_0^h AS(t)x\,dt = AS(0)x = Ax.$$

If $T(t)$ is a semigroup generated by A then for $t > 0$,

$$T'(t) \equiv \lim_{h\to 0} \frac{T(t+h) - T(t)}{h} = \lim_{h\to 0} \frac{T(h) - I}{h} T(t) = AT(t)$$

and $T(0) = I$. However, uniqueness applies because T and S both satisfy the same initial value problem and this yields $T(t) = S(t)$.

To show the power series equals $S(t)$ it suffices to show it satisfies the initial value problem. Using the mean value theorem,

$$\sum_{n=0}^{\infty} \frac{A^n((t+h)^n - t^n)}{n!} = \sum_{n=1}^{\infty} \frac{A^n(t+\theta_n(h))^{n-1}}{(n-1)!}$$

where $\theta_n(h) \in (0,h)$. Then taking a limit as $h \to 0$ and using the dominated convergence theorem, the limit of the difference quotient is

$$\sum_{n=1}^{\infty} \frac{A^n t^{n-1}}{(n-1)!} = A \sum_{n=1}^{\infty} \frac{A^{n-1} t^{n-1}}{(n-1)!} = A \sum_{n=0}^{\infty} \frac{(At)^n}{n!}$$

Thus $\sum_{n=0}^{\infty} \frac{(At)^n}{n!}$ satisfies the differential equation. It clearly satisfies the initial condition. Hence it equals $S(t)$. ■

Note that as a consequence of the above argument showing that T and S are the same, it follows that $T(t)A = AT(t)$ so one obtains that if the generator is a bounded linear operator, then the semigroup commutes with this operator.

When dealing with differential equations, one of the best tools is Gronwall's inequality. This is presented next.

Theorem 22.7.6 *Suppose u is nonnegative, continuous, and real valued and that*

$$u(t) \leq C + \int_0^t ku(s)\,ds,\ k \geq 0$$

Then $u(t) \leq Ce^{kt}$.

Proof: Let $w(t) \equiv \int_0^t ku(s)\,ds$. Then

$$w'(t) = ku(t) \leq kC + kw(t)$$

and so $w'(t) - kw(t) \leq kC$ which implies $\frac{d}{dt}\left(e^{-kt} w(t)\right) \leq kCe^{-kt}$. Therefore,

$$e^{-kt} w(t) \leq Ck \int_0^t e^{-ks} ds = Ck\left(\frac{1}{k} - \frac{1}{k} e^{-kt}\right)$$

so $w(t) \leq C\left(e^{kt} - 1\right)$. From the original inequality, $u(t) \leq C + w(t) \leq C + Ce^{kt} - C = Ce^{kt}$. ■

22.8 General Theory of Continuous Semigroups

Much more on semigroups is available in Yosida [59]. This is just an introduction to the subject.

22.8.1 Generators of Semigroups

Definition 22.8.1 *A strongly continuous semigroup defined on X,a Banach space is a function $S : [0,\infty) \to X$ which satisfies the following for all $x_0 \in X$.*

$$\begin{array}{rcl} S(t) & \in & \mathscr{L}(X,X), S(t+s) = S(t)S(s), \\ t & \to & S(t)x_0 \text{ is continuous, } \lim\limits_{t \to 0+} S(t)x_0 = x_0 \end{array}$$

Sometimes such a semigroup is said to be C_0. It is said to have the linear operator A as its generator if

$$D(A) \equiv \left\{ x : \lim_{h \to 0} \frac{S(h)x - x}{h} \text{ exists} \right\}$$

and for $x \in D(A)$, A is defined by

$$\lim_{h\to 0} \frac{S(h)x - x}{h} \equiv Ax$$

The assertion that $t \to S(t)x_0$ is continuous and that $S(t) \in \mathscr{L}(X,X)$ is not sufficient to say there is a bound on $\|S(t)\|$ for all $t \geq 0$. Also the assertion that for each $x_0, \lim_{t\to 0+} S(t)x_0 = x_0$ is not the same as saying that $S(t) \to I$ in $\mathscr{L}(X,X)$. It is a much weaker assertion. The next theorem gives information on the growth of $\|S(t)\|$. It turns out it has exponential growth. Thus $S(t)$ is a lot like e^t.

Lemma 22.8.2 *Let $M \equiv \sup\{\|S(t)\| : t \in [0,T]\}$. Then $M < \infty$.*

Proof: If this is not true, then there exists $t_n \in [0,T]$ such that $\|S(t_n)\| \geq n$. That is the operators $S(t_n)$ are not uniformly bounded. By the uniform boundedness principle, Theorem 21.1.9, there exists $x \in X$ such that $\|S(t_n)x\|$ is not bounded. However, this is impossible because it is given that $t \to S(t)x$ is continuous on $[0,T]$ and so $t \to \|S(t)x\|$ must achieve its maximum on this compact set. ■

Now here is the main result for growth of $\|S(t)\|$.

Theorem 22.8.3 *For M described in Lemma 22.8.2, there exists α such that*

$$\|S(t)\| \leq Me^{\alpha t}, t \geq 0$$

In fact, α can be chosen such that $M^{1/T} = e^{\alpha}$.

Proof: Let t be arbitrary. Then $t = mT + r(t)$ where $0 \leq r(t) < T$. Then by the semigroup property

$$\|S(t)\| = \|S(mT + r(t))\| = \|S(r(t))S(T)^m\| \leq M^{m+1}$$

Now $mT \leq t \leq mT + r(t) \leq (m+1)T$ and so $m \leq \frac{t}{T} \leq m+1$. Therefore,

$$\|S(t)\| \leq M^{(t/T)+1} = M\left(M^{1/T}\right)^t.$$

Let $M^{1/T} \equiv e^{\alpha}$ and then $\|S(t)\| \leq Me^{\alpha t}$ ■

Definition 22.8.4 *Let $S(t)$ be a continuous semigroup as described above. It is called a contraction semigroup if for all $t \geq 0, \|S(t)\| \leq 1$. It is called a bounded semigroup if there exists M such that for all $t \geq 0, \|S(t)\| \leq M$.*

Note that for $S(t)$ an arbitrary continuous semigroup satisfying $\|S(t)\| \leq Me^{\alpha t}$, It follows that the semigroup, $T(t) = e^{-\alpha t}S(t)$ is a bounded semigroup which satisfies $\|T(t)\| \leq M$.

The next proposition has to do with taking a Laplace transform of a semigroup. It suffices to let the integral be the usual Riemann integral for a function having values in a Banach space. You define it the same way as in single variable calculus in terms of limits of Riemann sums. Later, this will be generalized.

Proposition 22.8.5 *Given a continuous semigroup $S(t)$, its generator A exists and is a closed densely defined operator. Furthermore, for $\|S(t)\| \leq Me^{\alpha t}$ and $\lambda > \alpha$, $\lambda I - A$ is one to one and onto from $D(A)$ to X. Also $(\lambda I - A)^{-1}$ maps X onto $D(A)$ and is in $\mathscr{L}(X,X)$. Also for these values of $\lambda > \alpha, (\lambda I - A)^{-1} x = \int_0^\infty e^{-\lambda t} S(t) x dt$. For $\lambda > \alpha$, the following estimate holds.*

$$\left\| (\lambda I - A)^{-1} \right\| \leq \frac{M}{|\lambda - \alpha|} \tag{22.30}$$

Proof: First note $D(A) \neq \emptyset$. In fact $0 \in D(A)$. It follows from Theorem 22.8.3 that for all λ larger than α, one can define a Laplace transform, $R(\lambda)x \equiv \int_0^\infty e^{-\lambda t} S(t) x dt \in X$. The integral is the ordinary improper Riemann integral. Note that for $\lambda > \alpha$, $R(\lambda) \in \mathscr{L}(X,X)$ thanks to the estimates. Indeed, approximating with Riemann sums, to justify the details,

$$\|R(\lambda)x\| \leq \int_0^\infty e^{-\lambda t} \|S(t)x\| dt \leq \int_0^\infty M e^{-(\lambda - \alpha)t} dt \, \|x\| \leq \frac{M}{|\lambda - \alpha|} \|x\| \tag{22.31}$$

Claim 1: For $\lambda > \alpha, R(\lambda)x \in D(A)$ and $x = (\lambda I - A) R(\lambda) x$ so $R(\lambda)$ is a right inverse of $(\lambda I - A)$.

Proof of Claim 1: From the semigroup formula,

$$\frac{S(h)R(\lambda)x - R(\lambda)x}{h} = \frac{e^{h\lambda} \int_0^\infty e^{-\lambda(t+h)} S(t+h) x dt - \int_0^\infty e^{-\lambda t} S(t) x dt}{h} =$$

$$\frac{e^{h\lambda} \int_h^\infty e^{-\lambda t} S(t) x dt - \int_0^\infty e^{-\lambda t} S(t) x dt}{h} = \frac{\left(e^{h\lambda} - 1\right) R(\lambda) x - e^{\lambda h} \int_0^h e^{-\lambda t} S(t) x dt}{h}$$

Then it follows that the limit as $h \to 0$ exists and equals $\lambda R(\lambda)x - x$ which by definition of A is $A(R(\lambda)x)$. So by definition, $R(\lambda)x \in D(A)$ as claimed, and $\lambda I R(\lambda)x - A(R(\lambda)x) = x$ and so $x = (\lambda I - A) R(\lambda) x$. This shows **Claim 1.**

Claim 2: $D(A)$ is dense in X and for any $x \in X, \lim_{\lambda \to \infty} \lambda R(\lambda) x = x$.

Proof of Claim 2: Note that $\int_0^\infty \lambda e^{-\lambda t} dt = 1$ and so

$$\begin{aligned} \|\lambda R(\lambda)x - x\| &= \left\| \int_0^\infty \lambda e^{-\lambda t} S(t) x dt - x \right\| = \left\| \int_0^\infty \lambda e^{-\lambda t} (S(t)x - x) dt \right\| \\ &\leq \int_0^\infty \lambda e^{-\lambda t} \|(S(t)x - x)\| dt \end{aligned}$$

which from the estimates and standard approximate identity type arguments converges to 0 as follows: Let $\varepsilon > 0$ be given. Then choose δ such that $\|(S(t)x - x)\| < \varepsilon$ if $0 \leq t \leq \delta$. Then for λ large enough the second term in the following is no more than ε

$$\int_0^\infty \lambda e^{-\lambda t} \|(S(t)x - x)\| dt \leq \int_0^\delta \lambda e^{-\lambda t} \varepsilon + \int_\delta^\infty \lambda e^{-\lambda t} \left(Me^{\alpha t} + 1\right) \|x\| dt.$$

Thus for λ large enough, $\int_0^\infty \lambda e^{-\lambda t} \|(S(t)x - x)\| dt < 2\varepsilon$. This shows that $D(A)$ is dense in X and for any $x, \lim_{\lambda \to \infty} \lambda R(\lambda) x = x$. This proves **Claim 2.**

Claim 3: For $\lambda > \alpha$, $x = R(\lambda)(\lambda I - A)x$ for $x \in D(A)$ so $(\lambda I - A)$ is one to one and $R(\lambda)$ is a left inverse also. Thus $R(\lambda) = (\lambda I - A)^{-1}$ and from 22.31, estimate 22.30 holds.

Proof of Claim 3: If $x \in D(A)$, you could approximate with Riemann sums and pass to a limit and obtain the following for $\lambda > \alpha$.

$$\left\| R(\lambda) \left(\frac{S(h)x - x}{h} \right) - R(\lambda) Ax \right\| = \left\| \int_0^\infty e^{-\lambda t} S(t) \left(\frac{S(h)x - x}{h} - Ax \right) dt \right\|$$

$$\leq \int_0^\infty \left\| e^{-\lambda t} S(t) \right\| \left\| \frac{S(h)x - x}{h} - Ax \right\| dt$$

Then, passing to a limit as $h \to 0+$, this integrand converges uniformly to 0 so for all $\lambda > \alpha$,

$$\lim_{h\to 0} R(\lambda) \left(\frac{S(h)x - x}{h} \right) = R(\lambda) Ax \tag{22.32}$$

Also, $S(h)$ commutes with $R(\lambda)$. This follows from approximating with Riemann sums and taking a limit. Thus also

$$\lim_{h\to 0} R(\lambda) \left(\frac{S(h)x - x}{h} \right) = \lim_{h\to 0} \left(\frac{S(h)R(\lambda)x - R(\lambda)x}{h} \right) = AR(\lambda)x$$

so we have for $x \in D(A)$, $R(\lambda)Ax = AR(\lambda)x$. However, this implies

$$R(\lambda)(\lambda I - A)x = (\lambda I - A)R(\lambda)x = x$$

from **Claim 1.** Thus $R(\lambda)$ is a left inverse of $(\lambda I - A)$. Since $R(\lambda) = (\lambda I - A)^{-1}$, this shows the estimate 22.30 from 22.31. This proves **Claim 3.**

Why is A a closed operator? Suppose $x_n \to x$ where $x_n \in D(A)$ and that $Ax_n \to \xi$. I need to show that this implies that $x \in D(A)$ and that $Ax = \xi$. Thus $x_n \to x$ and for $\lambda > \alpha$, $(\lambda I - A)x_n \to \lambda x - \xi$. However, 22.30 shows that $(\lambda I - A)^{-1} = R(\lambda)$ is continuous and so

$$x_n \to (\lambda I - A)^{-1}(\lambda x - \xi) = x$$

It follows that $x \in D(A)$. Then doing $(\lambda I - A)$ to both sides of the equation, $\lambda x - \xi = \lambda x - Ax$ and so $Ax = \xi$ showing that A is a closed operator as claimed. ■

Definition 22.8.6 *The linear mapping for $\lambda > \alpha$ where $\|S(t)\| \leq Me^{\alpha t}$ given by $(\lambda I - A)^{-1} = R(\lambda)$ is called the resolvent.*

The following corollary is also very interesting.

Corollary 22.8.7 *Let $S(t)$ be a continuous semigroup and let A be its generator. Then for $0 < a < b < \infty$ and $x \in D(A)$, $S(b)x - S(a)x = \int_a^b S(t)Axdt$ and also for $t > 0$ you can take the derivative from the left,*

$$\lim_{h\to 0+} \frac{S(t)x - S(t-h)x}{h} = S(t)Ax$$

Proof: Letting $y^* \in X'$, you can take y^* inside the integral by approximating with Riemann sums. Thus

$$y^* \left(\int_a^b S(t)Axdt \right) = \int_a^b y^* \left(S(t) \lim_{h\to 0} \frac{S(h)x - x}{h} \right) dt$$

The difference quotients are bounded because they converge to Ax. Therefore, from the dominated convergence theorem and using the semigroup property,

$$y^* \left(\int_a^b S(t)Axdt \right) = \lim_{h\to 0} \int_a^b y^* \left(S(t) \frac{S(h)x - x}{h} \right) dt$$

$$= \lim_{h\to 0}\left(\frac{1}{h}\int_{a+h}^{b+h} y^* S(t)\,x\,dt - \frac{1}{h}\int_a^b y^* S(t)\,x\,dt\right)$$

$$= \lim_{h\to 0}\left(\frac{1}{h}\int_b^{b+h} y^* S(t)\,x\,dt - \frac{1}{h}\int_a^{a+h} y^* S(t)\,x\,dt\right) = y^*\left(S(b)\,x - S(a)\,x\right)$$

Since y^* is arbitrary, this proves the first part. Now from what was just shown, if $t > 0$ and h is small enough,

$$\frac{S(t)\,x - S(t-h)\,x}{h} = \frac{1}{h}\int_{t-h}^{t} S(s)\,Ax\,ds$$

which converges to $S(t)\,Ax$ as $h \to 0+$. ■

22.8.2 Hille Yosida Theorem

Given a closed densely defined operator, when is it the generator of a continuous semigroup? This is answered in the following theorem which is called the Hille Yosida theorem. It concerns the case of a bounded semigroup. However, if you have an arbitrary continuous semigroup, $S(t)$, then it was shown above that $S(t)\,e^{-\alpha t}$ is bounded for suitable α so the case discussed below is obtained.

Theorem 22.8.8 *Suppose A is a densely defined linear operator which has the property that for all $\lambda > 0$,*

$$(\lambda I - A)^{-1} \in \mathscr{L}(X,X)$$

which means that $\lambda I - A : D(A) \to X$ is one to one and onto with continuous inverse. Suppose also that for all $n \in \mathbb{N}$,

$$\left\|\left((\lambda I - A)^{-1}\right)^n\right\| \leq \frac{M}{\lambda^n}. \tag{22.33}$$

Then there exists a continuous semigroup $S(t)$ which has A as its generator and satisfies $\|S(t)\| \leq M$ and A is closed. In fact letting

$$S_\lambda(t) \equiv \exp\left(-\lambda + \lambda^2(\lambda I - A)^{-1}\right) \equiv \exp(A_\lambda)$$

it follows $\lim_{\lambda\to\infty} S_\lambda(t)\,x = S(t)\,x$ uniformly on finite intervals. Conversely, if A is the generator of $S(t)$, a bounded continuous semigroup having $\|S(t)\| \leq M$, then $(\lambda I - A)^{-1} \in \mathscr{L}(X,X)$ for all $\lambda > 0$ and 22.33 holds.

Proof: The condition 22.33 implies, that $\left\|(\lambda I - A)^{-1}\right\| \leq \frac{M}{\lambda}$.

Consider, for $\lambda > 0$, the operator which is defined on $D(A)$, $\lambda(\lambda I - A)^{-1} A$. On $D(A)$, this equals

$$-\lambda I + \lambda^2(\lambda I - A)^{-1} \tag{22.34}$$

because

$$\begin{aligned}(\lambda I - A)\,\lambda(\lambda I - A)^{-1} A &= \lambda A\\ (\lambda I - A)\left(-\lambda I + \lambda^2(\lambda I - A)^{-1}\right) &= -\lambda(\lambda I - A) + \lambda^2 = \lambda A\end{aligned}$$

and, by assumption, $(\lambda I - A)$ is one to one. From the second line of 22.34, the operator $-\lambda I + \lambda^2 (\lambda I - A)^{-1}$ makes sense on all of X not just on $D(A)$. Also

$$\left(-\lambda I + \lambda^2 (\lambda I - A)^{-1}\right)(\lambda I - A) = -\lambda (\lambda I - A) + \lambda^2 I = \lambda A$$

$$\lambda A (\lambda I - A)^{-1} (\lambda I - A) = \lambda A$$

so, since $(\lambda I - A)$ is onto, it follows that on X,

$$-\lambda I + \lambda^2 (\lambda I - A)^{-1} = A\lambda (\lambda I - A)^{-1} \equiv A_\lambda$$

Denote this as A_λ to save notation. Thus on $D(A)$,

$$\lambda A (\lambda I - A)^{-1} = \lambda (\lambda I - A)^{-1} A = A_\lambda$$

although the $\lambda (\lambda I - A)^{-1} A$ only makes sense on $D(A)$.This is summarized next.

Lemma 22.8.9 *There is a bounded linear operator given for $\lambda > 0$ by*

$$-\lambda I + \lambda^2 (\lambda I - A)^{-1} = \lambda A (\lambda I - A)^{-1} \equiv A_\lambda$$

On $D(A)$, $A_\lambda = \lambda (\lambda I - A)^{-1} A$. Also, for all $x \in X$,

$$\lim_{\lambda \to \infty} \lambda (\lambda I - A)^{-1} x - x = 0. \tag{22.35}$$

Replacing x with Ax, it follows that for all $x \in D(A)$,

$$\lim_{\lambda \to \infty} A_\lambda x = Ax. \tag{22.36}$$

Proof: First assume $x \in D(A)$

$$\begin{aligned} & \left\| \lambda (\lambda I - A)^{-1} x - x \right\| = \left\| (\lambda I - A)^{-1} (\lambda x - (\lambda I - A) x) \right\| \\ = \quad & \left\| (\lambda I - A)^{-1} Ax \right\| \leq \frac{M}{\lambda} \|Ax\| \end{aligned} \tag{22.37}$$

which converges to 0 as $\lambda \to \infty$.

Now let x be general and let $\hat{x} \in D(A)$. From 22.33, $\left| \lambda (\lambda I - A)^{-1} \right| \leq M$. Then

$$\begin{aligned} \left\| \lambda (\lambda I - A)^{-1} x - x \right\| \quad \leq \quad & \left\| \lambda (\lambda I - A)^{-1} x - \lambda (\lambda I - A)^{-1} \hat{x} \right\| \\ & + \left\| \lambda (\lambda I - A)^{-1} \hat{x} - \hat{x} \right\| + \|\hat{x} - x\| \end{aligned}$$

Let $\hat{x}$ be close enough to x that the first and last terms on the right added together are less than $\varepsilon/2$. Then whenever λ is large enough, the first part of the argument shows that the middle term is no more than $\varepsilon/2$. This verifies 22.35, 22.36. ■

Now from Corollary 22.7.5, there exists an approximate continuous semigroup $S_\lambda(t)$ generated by A_λ which is the solution to

$$S_\lambda'(t) = A_\lambda S_\lambda(t), S_\lambda(0) = I \tag{22.38}$$

In terms of power series,

$$S_\lambda(t) \equiv e^{-\lambda t}\sum_{k=0}^{\infty}\frac{t^k\left(\lambda^2(\lambda I-A)^{-1}\right)^k}{k!} = e^{t\left(-\lambda I+\lambda^2(\lambda I-A)^{-1}\right)} \tag{22.39}$$

Thus, by assumption 22.33 and triangle inequality,

$$\|S_\lambda(t)\| \le e^{-\lambda t}\sum_{k=0}^{\infty}\frac{t^k}{k!}\lambda^{2k}\frac{M}{\lambda^k} = e^{-\lambda t}Me^{\lambda t} = M \tag{22.40}$$

Next is an easy observation about operators commuting.

Lemma 22.8.10 *For $\lambda,\mu>0$, $(\lambda I-A)^{-1}$ and $(\mu I-A)^{-1}$ commute.*

Proof: Suppose

$$y = (\mu I-A)^{-1}(\lambda I-A)^{-1}x \tag{22.41}$$
$$z = (\lambda I-A)^{-1}(\mu I-A)^{-1}x \tag{22.42}$$

I need to show $y=z$. This follows from the observation that

$$(\lambda I-A)(\mu I-A)y = (\mu I-A)(\lambda I-A)y = (\mu I-A)(\lambda I-A)z = x$$

■

It follows from the description of $S_\lambda(t)$ in terms of a power series that $S_\lambda(t)$ and $S_\mu(s)$ commute and also A_λ commutes with $S_\mu(t)$ for any t. Indeed, the absolute convergence of the series 22.39 means we can use the Cauchy product to compute the product of these two series and see $S_\lambda(t), S_\mu(t)$ commute. One could also exploit uniqueness and the theory of ordinary differential equations to verify this. I will use this fact in what follows whenever needed.

I want to show that for each $x\in D(A)$,

$$\lim_{\lambda\to\infty} S_\lambda(t)x \equiv S(t)x$$

where $S(t)$ is the desired semigroup. Let $x\in D(A)$. Then

$$S_\mu(t)x - S_\lambda(t)x = \int_0^t \frac{d}{dr}\left(S_\lambda(t-r)S_\mu(r)\right)xdr$$

$$= \int_0^t \left(-S'_\lambda(t-r)S_\mu(r) + S_\lambda(t-r)S'_\mu(r)\right)xdr$$
$$= \int_0^t \left(S_\lambda(t-r)S_\mu(r)A_\lambda - S_\mu(r)S_\lambda(t-r)A_\mu\right)xdr$$
$$= \int_0^t S_\lambda(t-r)S_\mu(r)\left(A_\mu x - A_\lambda x\right)dr$$

It follows that

$$\left\|S_\mu(t)x - S_\lambda(t)x\right\| \le \int_0^t \left\|S_\lambda(t-r)S_\mu(r)\left(A_\mu x - A_\lambda x\right)\right\|dr$$

$$\leq M^2 t \left\|A_\mu x - A_\lambda x\right\| \leq M^2 t \left(\left\|A_\mu x - Ax\right\| + \left\|Ax - A_\lambda x\right\|\right)$$

Now by Lemma 22.8.9, the right side converges uniformly to 0 in $t \in [0,T]$ an arbitrary finite interval. Denote that to which it converges $S(t)x$. Therefore, $t \to S(t)x$ is continuous for each $x \in D(A)$ and also from 22.40,

$$\|S(t)x\| = \lim_{\lambda \to \infty} \|S_\lambda(t)x\| \leq M\|x\|$$

so that $S(t)$ can be extended uniquely to a continuous linear map, still called $S(t)$ defined on all of X which also satisfies $\|S(t)\| \leq M$ since $D(A)$ is dense in X. The uniform convergence on $[0,T]$ implies $t \to S(t)$ is continuous.

It remains to verify that A generates $S(t)$ and for all x, $\lim_{t\to 0+} S(t)x - x = 0$. From the above,

$$S_\lambda(t)x = x + \int_0^t S_\lambda(s) A_\lambda x ds \tag{22.43}$$

and so $\lim_{t\to 0+} \|S_\lambda(t)x - x\| = 0$. By the uniform convergence just shown, there exists λ large enough that for all $t \in [0,\delta]$, $\|S(t)x - S_\lambda(t)x\| < \varepsilon$. Then

$$\begin{aligned} \limsup_{t\to 0+} \|S(t)x - x\| &\leq \limsup_{t\to 0+} \left(\|S(t)x - S_\lambda(t)x\| + \|S_\lambda(t)x - x\|\right) \\ &\leq \limsup_{t\to 0+} \left(\varepsilon + \|S_\lambda(t)x - x\|\right) \leq \varepsilon \end{aligned}$$

It follows $\lim_{t\to 0+} S(t)x = x$ because ε is arbitrary.

Next, $\lim_{\lambda\to\infty} A_\lambda x = Ax$ for all $x \in D(A)$ by Lemma 22.8.9. Therefore, passing to the limit in 22.43 yields from the uniform convergence

$$S(t)x = x + \int_0^t S(s) Ax ds$$

and by continuity of $s \to S(s)Ax$, it follows

$$\lim_{h\to 0+} \frac{S(h)x - x}{h} = \lim_{h\to 0} \frac{1}{h} \int_0^h S(s) Ax ds = Ax$$

Thus letting B denote the generator of $S(t)$, $D(A) \subseteq D(B)$ and $A = B$ on $D(A)$. It only remains to verify $D(A) = D(B)$.

To do this, let $\lambda > 0$ and consider the following where $y \in X$ is arbitrary.

$$(\lambda I - B)^{-1} y = (\lambda I - B)^{-1} \left((\lambda I - A)(\lambda I - A)^{-1} y\right)$$

Now $(\lambda I - A)^{-1} y \in D(A) \subseteq D(B)$ and $A = B$ on $D(A)$ and so

$$(\lambda I - A)(\lambda I - A)^{-1} y = (\lambda I - B)(\lambda I - A)^{-1} y$$

which implies,

$$(\lambda I - B)^{-1} y =$$

$$(\lambda I - B)^{-1} \left((\lambda I - B)(\lambda I - A)^{-1} y\right) = (\lambda I - A)^{-1} y$$

Recall from Proposition 22.8.5, an arbitrary element of $D(B)$ is of the form $(\lambda I - B)^{-1} y$ and this has shown every such vector is in $D(A)$, in fact it equals $(\lambda I - A)^{-1} y$. Hence $D(B) \subseteq D(A)$ which shows A generates $S(t)$ and this proves the first half of the theorem.

Next suppose A is the generator of a semigroup $S(t)$ having $\|S(t)\| \leq M$. Then by Proposition 22.8.5 for all $\lambda > 0, (\lambda I - A)$ is onto and $(\lambda I - A)^{-1} = \int_0^\infty e^{-\lambda t} S(t)\, dt$. Thus,

$$\begin{aligned} & \left\| \left((\lambda I - A)^{-1} \right)^n \right\| \\ = & \left\| \int_0^\infty \cdots \int_0^\infty e^{-\lambda(t_1+\cdots+t_n)} S(t_1+\cdots+t_n)\, dt_1 \cdots dt_n \right\| \\ \leq & \int_0^\infty \cdots \int_0^\infty e^{-\lambda(t_1+\cdots+t_n)} M dt_1 \cdots dt_n = \frac{M}{\lambda^n}. \blacksquare \end{aligned}$$

22.8.3 An Evolution Equation

When Λ generates a continuous semigroup, one can consider a very interesting theorem about evolution equations of the form $y' - \Lambda y = g(t)$ provided $t \to g(t)$ is C^1.

Theorem 22.8.11 *Let Λ be the generator of $S(t)$, a continuous semigroup on X, a Banach space and let $t \to g(t)$ be in $C^1(0,\infty;X)$. Then there exists a unique solution to the initial value problem $y' = \Lambda y + g$, $y(0) = y_0 \in D(\Lambda)$ and it is given by*

$$y(t) = S(t) y_0 + \int_0^t S(t-s) g(s)\, ds. \tag{22.44}$$

This solution is continuous having continuous derivative and has values in $D(\Lambda)$.

Proof: First I show the following claim.
Claim: For $t > 0$, $\int_0^t S(t-s) g(s)\, ds \in D(\Lambda)$ and

$$\Lambda \left(\int_0^t S(t-s) g(s)\, ds \right) = S(t) g(0) - g(t) + \int_0^t S(t-s) g'(s)\, ds$$

Proof of the claim:

$$\begin{aligned} & \frac{1}{h} \left(S(h) \int_0^t S(t-s) g(s)\, ds - \int_0^t S(t-s) g(s)\, ds \right) \\ = & \frac{1}{h} \left(\int_0^t S(t-s+h) g(s)\, ds - \int_0^t S(t-s) g(s)\, ds \right) \\ = & \frac{1}{h} \left(\int_{-h}^{t-h} S(t-s) g(s+h)\, ds - \int_0^t S(t-s) g(s)\, ds \right) \\ = & \frac{1}{h} \int_{-h}^0 S(t-s) g(s+h)\, ds + \int_0^{t-h} S(t-s) \frac{g(s+h) - g(s)}{h} \\ & - \frac{1}{h} \int_{t-h}^t S(t-s) g(s)\, ds \end{aligned}$$

Using the estimate in Theorem 22.8.3 on Page 625, the triangle inequality and the uniform convergence of the integrands, the limit as $h \to 0$ of the above equals

$$S(t) g(0) - g(t) + \int_0^t S(t-s) g'(s)\, ds$$

which proves the claim since the limit exists and is therefore, $\Lambda\left(\int_0^t S(t-s)g(s)\,ds\right)$.

Since $y_0 \in D(\Lambda)$,

$$\begin{aligned} S(t)\Lambda y_0 &= S(t)\lim_{h\to 0}\frac{S(h)y_0 - y_0}{h} = \lim_{h\to 0}\frac{S(t+h)-S(t)}{h}y_0 \\ &= \lim_{h\to 0+}\frac{S(h)S(t)y_0 - S(t)y_0}{h} \equiv \Lambda S(t)y_0 \end{aligned} \tag{22.45}$$

This is because the limit exists and so it is by definition the right side. So $S(t)y_0 \in D(\Lambda)$. Now consider 22.44.

$$\frac{y(t+h)-y(t)}{h} = \frac{S(t+h)-S(t)}{h}y_0 + \frac{1}{h}\left(\int_0^{t+h} S(t-s+h)g(s)\,ds - \int_0^t S(t-s)g(s)\,ds\right)$$

$$\begin{aligned} &= \frac{S(t+h)-S(t)}{h}y_0 + \frac{1}{h}\int_t^{t+h} S(t-s+h)g(s)\,ds \\ &\quad + \frac{1}{h}\left(S(h)\int_0^t S(t-s)g(s)\,ds - \int_0^t S(t-s)g(s)\,ds\right) \end{aligned}$$

From the claim and 22.45, the limit of the right side is

$$\begin{aligned} &\Lambda S(t)y_0 + g(t) + \Lambda\left(\int_0^t S(t-s)g(s)\,ds\right) \\ &= \Lambda\left(S(t)y_0 + \int_0^t S(t-s)g(s)\,ds\right) + g(t) \end{aligned}$$

Hence $y'(t) = \Lambda y(t) + g(t)$ and from the formula, y' is continuous since by the claim and 22.45 it also equals

$$S(t)\Lambda y_0 + g(t) + S(t)g(0) - g(t) + \int_0^t S(t-s)g'(s)\,ds$$

which is continuous. The claim and 22.45 also shows $y(t) \in D(\Lambda)$. This proves the existence part of the lemma.

It remains to prove the uniqueness part. It suffices to show that if

$$y' - \Lambda y = 0,\ y(0) = 0$$

and y is C^1 having values in $D(\Lambda)$, then $y = 0$. Suppose then that y is this way. Letting $0 < s < t$,

$$\frac{d}{ds}(S(t-s)y(s)) \equiv$$

$$\lim_{h\to 0} S(t-s-h)\frac{y(s+h)-y(s)}{h} - \frac{S(t-s)y(s) - S(t-s-h)y(s)}{h}$$

provided the limit exists. Since y' exists and $y(s) \in D(\Lambda)$, this equals

$$S(t-s)y'(s) - S(t-s)\Lambda y(s) = 0.$$

Let $y^* \in X'$. This has shown that on the open interval $(0,t)$, $s \to y^*(S(t-s)y(s))$ has a derivative equal to 0. Also from continuity of S and y, this function is continuous on $[0,t]$. Therefore, it is constant on $[0,t]$ by the mean value theorem. At $s=0$, this function equals 0. Therefore, it equals 0 on $[0,t]$. Thus for fixed $s>0$ and letting $t>s, y^*(S(t-s)y(s))=0$. Now let t decrease toward s. Then $y^*(y(s))=0$ and since y^* was arbitrary, it follows $y(s)=0$. ■

22.8.4 Adjoints for Closed Operators, Hilbert Space

In Hilbert space, there are some special things which are true.

Definition 22.8.12 *Let A be a densely defined closed operator on H a real Hilbert space. Then A^* is defined as follows.*

$$D(A^*) \equiv \{y \in H : |(Ax,y)| \leq C|x|\}$$

Then since $D(A)$ is dense, there exists a unique element of H denoted by A^*y such that

$$(Ax,y) = (x,A^*y)$$

for all $x \in D(A)$.

Lemma 22.8.13 *Let A be closed and densely defined on $D(H) \subseteq H$, a Hilbert space. Then A^* is also closed and densely defined. Also $(A^*)^* = A$. If $(\lambda I - A)^{-1} \in \mathscr{L}(H,H)$, then $(\lambda I - A^*)^{-1} \in \mathscr{L}(H,H)$ and $\left(\left((\lambda I - A)^{-1}\right)^n\right)^* = \left((\lambda I - A^*)^{-1}\right)^n$.*

Proof: Denote by $[x,y]$ an ordered pair in $H \times H$. Define $\tau : H \times H \to H \times H$ by

$$\tau[x,y] \equiv [-y,x]$$

Then the definition of adjoint implies that for $\mathscr{G}(B)$ equal to the graph of B,

$$\mathscr{G}(A^*) = (\tau\mathscr{G}(A))^{\perp} \tag{22.46}$$

In this notation the inner product on $H \times H$ with respect to which $\perp$ is defined is given by

$$([x,y],[a,b]) \equiv (x,a) + (y,b).$$

Here is why this is so. For $[x,A^*x] \in \mathscr{G}(A^*)$ it follows that for all $y \in D(A)$

$$([x,A^*x],[-Ay,y]) = -(Ay,x) + (y,A^*x) = 0$$

and so $[x,A^*x] \in (\tau\mathscr{G}(A))^{\perp}$ which shows $\mathscr{G}(A^*) \subseteq (\tau\mathscr{G}(A))^{\perp}$. To obtain the other inclusion, let $[a,b] \in (\tau\mathscr{G}(A))^{\perp}$. This means that for all $x \in D(A)$,

$$([a,b],[-Ax,x]) = 0.$$

In other words, for all $x \in D(A)$, $(Ax,a) = (x,b)$ and so $|(Ax,a)| \leq C|x|$ for all $x \in D(A)$ which shows $a \in D(A^*)$ and $(x,A^*a) = (x,b)$ for all $x \in D(A)$. Therefore, since $D(A)$ is dense, it follows $b = A^*a$ and so $[a,b] \in \mathscr{G}(A^*)$. This shows the other inclusion.

Note that if V is any subspace of the Hilbert space $H\times H$, $\left(V^{\perp}\right)^{\perp}=\overline{V}$ and $S^{\perp}$ is always a closed subspace. Also τ and $\perp$ commute. The reason for this is that $[x,y]\in(\tau V)^{\perp}$ means that $(x,-b)+(y,a)=0.$ for all $[a,b]\in V$ and $[x,y]\in\tau\left(V^{\perp}\right)$ means $[-y,x]\in V^{\perp}$ so for all $[a,b]\in V$, $(-y,a)+(x,b)=0.$which says the same thing. It is also clear that $\tau\circ\tau$ has the effect of multiplication by -1.

It follows from the above description of the graph of A^{*} that even if $\mathscr{G}(A)$ were not closed it would still be the case that $\mathscr{G}(A^{*})$ is closed.

Why is $D(A^{*})$ dense? Suppose $z\in D(A^{*})^{\perp}$. Then for all $y\in D(A^{*})$ so that $[y,Ay]\in\mathscr{G}(A^{*})$, it follows $[z,0]\in\mathscr{G}(A^{*})^{\perp}=\left((\tau\mathscr{G}(A))^{\perp}\right)^{\perp}=\tau\mathscr{G}(A)$ but this implies $[0,z]\in-\mathscr{G}(A)$. and so $z=-A0=0$. Thus $D(A^{*})$ must be dense since there is no nonzero vector in $D(A^{*})^{\perp}$.

Since A is a closed operator, meaning $\mathscr{G}(A)$ is closed in $H\times H$, it follows from the above formula that

$$\begin{aligned}\mathscr{G}\left((A^{*})^{*}\right) &= \left(\tau\left((\tau\mathscr{G}(A))^{\perp}\right)\right)^{\perp}=\left(\tau(\tau\mathscr{G}(A))^{\perp}\right)^{\perp}\\ &= \left((-\mathscr{G}(A))^{\perp}\right)^{\perp}=\left(\mathscr{G}(A)^{\perp}\right)^{\perp}=\mathscr{G}(A)\end{aligned}$$

and so $(A^{*})^{*}=A$.

Now consider the final claim. First let $y\in D(A^{*})=D(\lambda I-A^{*})$. Then letting $x\in H$ be arbitrary,

$$\left(x,\left((\lambda I-A)(\lambda I-A)^{-1}\right)^{*}y\right)$$

$$\left((\lambda I-A)(\lambda I-A)^{-1}x,y\right)=\left(x,\left((\lambda I-A)^{-1}\right)^{*}(\lambda I-A^{*})y\right)$$

Thus

$$\left((\lambda I-A)(\lambda I-A)^{-1}\right)^{*}=I=\left((\lambda I-A)^{-1}\right)^{*}(\lambda I-A^{*}) \tag{22.47}$$

on $D(A^{*})$. Next let $x\in D(A)=D(\lambda I-A)$ and $y\in H$ arbitrary.

$$(x,y)=\left((\lambda I-A)^{-1}(\lambda I-A)x,y\right)=\left((\lambda I-A)x,\left((\lambda I-A)^{-1}\right)^{*}y\right)$$

Now it follows $\left|\left((\lambda I-A)x,\left((\lambda I-A)^{-1}\right)^{*}y\right)\right|\leq|y||x|$ for any $x\in D(A)$ and so

$$\left((\lambda I-A)^{-1}\right)^{*}y\in D(A^{*})$$

Hence

$$(x,y)=\left(x,(\lambda I-A^{*})\left((\lambda I-A)^{-1}\right)^{*}y\right).$$

Since $x\in D(A)$ is arbitrary and $D(A)$ is dense, it follows

$$(\lambda I-A^{*})\left((\lambda I-A)^{-1}\right)^{*}=I \tag{22.48}$$

From 22.47 and 22.48 it follows $(\lambda I-A^{*})^{-1}=\left((\lambda I-A)^{-1}\right)^{*}$ and $(\lambda I-A^{*})$ is one to one and onto with continuous inverse. Finally, from the above,

$$\left((\lambda I-A^{*})^{-1}\right)^{n}=\left(\left((\lambda I-A)^{-1}\right)^{*}\right)^{n}=\left(\left((\lambda I-A)^{-1}\right)^{n}\right)^{*}.\blacksquare$$

With this preparation, here is an interesting result about the adjoint of the generator of a continuous bounded semigroup. I found this in Balakrishnan [5].

Theorem 22.8.14 *Suppose A is a densely defined closed operator which generates a continuous semigroup, $S(t)$. Then A^* is also a closed densely defined operator which generates $S^*(t)$ and $S^*(t)$ is also a continuous semigroup.*

Proof: First suppose $S(t)$ is also a bounded semigroup, $\|S(t)\| \leq M$. From Lemma 22.8.13 A^* is closed and densely defined. It follows from the Hille Yosida theorem, Theorem 22.8.8 that

$$\left|\left((\lambda I - A)^{-1}\right)^n\right| \leq \frac{M}{\lambda^n}$$

From Lemma 22.8.13 and the fact the adjoint of a bounded linear operator preserves the norm,

$$\begin{aligned} \frac{M}{\lambda^n} &\geq \left|\left(\left((\lambda I - A)^{-1}\right)^n\right)^*\right| = \left|\left(\left((\lambda I - A)^{-1}\right)^*\right)^n\right| \\ &= \left|\left((\lambda I - A^*)^{-1}\right)^n\right| \end{aligned}$$

and so by Theorem 22.8.8 again it follows A^* generates a continuous semigroup, $T(t)$ which satisfies $\|T(t)\| \leq M$. I need to identify $T(t)$ with $S^*(t)$. However, from the proof of Theorem 22.8.8 and Lemma 22.8.13, it follows that for $x \in D(A^*)$ and a suitable sequence $\{\lambda_n\}$,

$$(T(t)x, y) = \left(\lim_{n\to\infty} e^{-\lambda_n t} \sum_{k=0}^{\infty} \frac{t^k \left(\lambda_n^2 (\lambda_n I - A^*)^{-1}\right)^k}{k!} x, y\right)$$

$$\begin{aligned} &= \lim_{n\to\infty} \left(e^{-\lambda_n t} \sum_{k=0}^{\infty} \frac{t^k \left(\left(\lambda_n^2 (\lambda_n I - A)^{-1}\right)^k\right)^*}{k!} x, y\right) \\ &= \lim_{n\to\infty} \left(x, e^{-\lambda_n t} \left(\sum_{k=0}^{\infty} \frac{t^k \left(\left(\lambda_n^2 (\lambda_n I - A)^{-1}\right)^k\right)}{k!}\right) y\right) \\ &= (x, S(t)y) = (S^*(t)x, y). \end{aligned}$$

Therefore, since y is arbitrary, $S^*(t) = T(t)$ on $x \in D(A^*)$ a dense set and this shows the two are equal. This proves the proposition in the case where $S(t)$ is also bounded.

Next only assume $S(t)$ is a continuous semigroup. Then by Proposition 22.8.5 there exists $\alpha > 0$ such that

$$\|S(t)\| \leq Me^{\alpha t}.$$

Then consider the operator $-\alpha I + A$ and the bounded semigroup $e^{-\alpha t} S(t)$. For $x \in D(A)$

$$\begin{aligned} \lim_{h\to 0+} \frac{e^{-\alpha h} S(h)x - x}{h} &= \lim_{h\to 0+} \left(e^{-\alpha h} \frac{S(h)x - x}{h} + \frac{e^{-\alpha h} - 1}{h} x\right) \\ &= -\alpha x + Ax \end{aligned}$$

Thus $-\alpha I+A$ generates $e^{-\alpha t}S(t)$ and it follows from the first part that $-\alpha I+A^*$ generates $e^{-\alpha t}S^*(t)$. Thus

$$\begin{aligned} -\alpha x+A^*x &= \lim_{h\to 0+}\frac{e^{-\alpha h}S^*(h)x-x}{h} \\ &= \lim_{h\to 0+}\left(e^{-\alpha h}\frac{S^*(h)x-x}{h}+\frac{e^{-\alpha h}-1}{h}x\right) \\ &= -\alpha x+\lim_{h\to 0+}\frac{S^*(h)x-x}{h} \end{aligned}$$

showing that A^* generates $S^*(t)$. It follows from Proposition 22.8.5 that A^* is closed and densely defined. It is obvious $S^*(t)$ is a semigroup. Why is it continuous? This also follows from the first part of the argument which establishes that $e^{-\alpha t}S^*(t)$ is continuous. This proves the theorem.

22.8.5 Adjoints, Reflexive Banach Space

Here the adjoint of a generator of a semigroup is considered. I will show that the adjoint of the generator generates the adjoint of the semigroup in a reflexive Banach space. This is about as far as you can go although a general but less satisfactory result is given in Yosida [59].

Definition 22.8.15 *Let A be a densely defined closed operator on H a real Banach space. Then A^* is defined as follows.*

$$D(A^*)\equiv\left\{y^*\in H' : |y^*(Ax)|\le C\,||x||\ \textit{for all}\ x\in D(A)\right\}$$

Then since $D(A)$ is dense, there exists a unique element of H' denoted by A^*y such that

$$A^*(y^*)(x)=y^*(Ax)$$

for all $x\in D(A)$.

Lemma 22.8.16 *Let A be closed and densely defined on $D(A)\subseteq H$, a Banach space. Then A^* is also closed and densely defined. Also $(A^*)^*=A$. In addition to this, if*

$$(\lambda I-A)^{-1}\in\mathscr{L}(H,H),$$

then $(\lambda I-A^*)^{-1}\in\mathscr{L}(H',H')$ and

$$\left(\left((\lambda I-A)^{-1}\right)^n\right)^*=\left((\lambda I-A^*)^{-1}\right)^n$$

Proof: Denote by $[x,y]$ an ordered pair in $H\times H$. Define $\tau : H\times H\to H\times H$ by

$$\tau[x,y]\equiv[-y,x]$$

A similar notation will apply to $H'\times H'$. Then the definition of adjoint implies that for $\mathscr{G}(B)$ equal to the graph of B,

$$\mathscr{G}(A^*)=(\tau\mathscr{G}(A))^{\perp} \tag{22.49}$$

For $S \subseteq H \times H$, define $S^{\perp}$ by

$$\left\{[a^*, b^*] \in H' \times H' : a^*(x) + b^*(y) = 0 \text{ for all } [x,y] \in S\right\}$$

If $S \subseteq H' \times H'$ a similar definition holds.

$$\{[x,y] \in H \times H : a^*(x) + b^*(y) = 0 \text{ for all } [a^*, b^*] \in S\}$$

Here is why 22.49 is so. For $[x^*, A^*x^*] \in \mathcal{G}(A^*)$ it follows that for all $y \in D(A)$

$$x^*(Ay) = A^*x^*(y)$$

and so for all $[y, Ay] \in \mathcal{G}(A)$,

$$-x^*(Ay) + A^*x^*(y) = 0$$

which is what it means to say $[x^*, A^*x^*] \in (\tau\mathcal{G}(A))^{\perp}$. This shows

$$\mathcal{G}(A^*) \subseteq (\tau\mathcal{G}(A))^{\perp}$$

To obtain the other inclusion, let $[a^*, b^*] \in (\tau\mathcal{G}(A))^{\perp}$. This means that for all $[x, Ax] \in \mathcal{G}(A)$,

$$-a^*(Ax) + b^*(x) = 0$$

In other words, for all $x \in D(A)$,

$$|a^*(Ax)| \leq \|b^*\| \|x\|$$

which means by definition, $a^* \in D(A^*)$ and $A^*a^* = b^*$. Thus $[a^*, b^*] \in \mathcal{G}(A^*)$.This shows the other inclusion.

Note that if V is any subspace of $H \times H$, $\left(V^{\perp}\right)^{\perp} = \overline{V}$. and $S^{\perp}$ is always a closed subspace. Also τ and $\perp$ commute. The reason for this is that $[x^*, y^*] \in (\tau V)^{\perp}$ means that

$$-x^*(b) + y^*(a) = 0$$

for all $[a,b] \in V$ and $[x^*, y^*] \in \tau\left(V^{\perp}\right)$ means $[-y^*, x^*] \in -\left(V^{\perp}\right) = V^{\perp}$ so for all $[a,b] \in V$,

$$-y^*(a) + x^*(b) = 0$$

which says the same thing. It is also clear that $\tau \circ \tau$ has the effect of multiplication by -1. If $V \subseteq H' \times H'$, the argument for commuting $\perp$ and τ is similar.

It follows from the above description of the graph of A^* that even if $\mathcal{G}(A)$ were not closed it would still be the case that $\mathcal{G}(A^*)$ is closed.

Why is $D(A^*)$ dense? If it is not dense, then by a typical application of the Hahn Banach theorem, there exists $y^{**} \in H''$ such that $y^{**}(D(A^*)) = 0$ but $y^{**} \neq 0$. Since H is reflexive, there exists $y \in H$ such that $x^*(y) = 0$ for all $x^* \in D(A^*)$. Thus

$$[y, 0] \in \mathcal{G}(A^*)^{\perp} = \left((\tau\mathcal{G}(A))^{\perp}\right)^{\perp} = \tau\mathcal{G}(A)$$

and so $[0, y] \in \mathcal{G}(A)$ which means $y = A0 = 0$, a contradiction. Thus $D(A^*)$ is indeed dense. Note this is where it was important to assume the space is reflexive. If you consider

$C([0,1])$ it is not dense in $L^\infty([0,1])$ but if $f \in L^1([0,1])$ satisfies $\int_0^1 fgdm = 0$ for all $g \in C([0,1])$, then $f = 0$. Hence there is no nonzero $f \in C([0,1])^\perp$.

Since A is a closed operator, meaning $\mathscr{G}(A)$ is closed in $H \times H$, it follows from the above formula that

$$\begin{aligned} \mathscr{G}\left((A^*)^*\right) &= \left(\tau\left((\tau\mathscr{G}(A))^\perp\right)\right)^\perp = \left(\tau\left(\tau\mathscr{G}(A)\right)^\perp\right)^\perp \\ &= \left((-\mathscr{G}(A))^\perp\right)^\perp = \left(\mathscr{G}(A)^\perp\right)^\perp = \mathscr{G}(A) \end{aligned}$$

and so $(A^*)^* = A$.

Now consider the final claim. First let $y^* \in D(A^*) = D(\lambda I - A^*)$. Then letting $x \in H$ be arbitrary,

$$y^*(x) = \left((\lambda I - A)(\lambda I - A)^{-1}\right)^* y^*(x) = y^*\left((\lambda I - A)(\lambda I - A)^{-1} x\right)$$

Since $y^* \in D(A^*)$ and $(\lambda I - A)^{-1} x \in D(A)$, this equals $(\lambda I - A)^* y^*\left((\lambda I - A)^{-1} x\right)$. Now by definition, this equals $\left((\lambda I - A)^{-1}\right)^* (\lambda I - A)^* y^*(x)$. It follows that for $y^* \in D(A^*)$,

$$\left((\lambda I - A)^{-1}\right)^* (\lambda I - A)^* y^* = \left((\lambda I - A)^{-1}\right)^* (\lambda I - A^*) y^* = y^* \tag{22.50}$$

Next let $y^* \in H'$ be arbitrary and $x \in D(A)$

$$\begin{aligned} y^*(x) &= y^*\left((\lambda I - A)^{-1}(\lambda I - A)x\right) = \left((\lambda I - A)^{-1}\right)^* y^*\left((\lambda I - A)x\right) \\ &= (\lambda I - A)^*\left((\lambda I - A)^{-1}\right)^* y^*(x) \end{aligned}$$

In going from the second to the third line, the first line shows $\left((\lambda I - A)^{-1}\right)^* y^* \in D(A^*)$ and so the third line follows. Since $D(A)$ is dense, it follows

$$(\lambda I - A^*)\left((\lambda I - A)^{-1}\right)^* = I \tag{22.51}$$

Then 22.50 and 22.51 show $\lambda I - A^*$ is one to one and onto from $D(A^*)$ to H' and

$$(\lambda I - A^*)^{-1} = \left((\lambda I - A)^{-1}\right)^*.$$

Finally, from the above, $\left((\lambda I - A^*)^{-1}\right)^n = \left(\left((\lambda I - A)^{-1}\right)^*\right)^n = \left(\left((\lambda I - A)^{-1}\right)^n\right)^*$. This proves the lemma.

With this preparation, here is an interesting result about the adjoint of the generator of a continuous bounded semigroup.

Theorem 22.8.17 *Suppose A is a densely defined closed operator which generates a continuous semigroup, $S(t)$. Then A^* is also a closed densely defined operator which generates $S^*(t)$ and $S^*(t)$ is also a continuous semigroup.*

Proof: First suppose $S(t)$ is also a bounded semigroup, $\|S(t)\| \leq M$. From Lemma 22.8.16 A^* is closed and densely defined. It follows from the Hille Yosida theorem, Theorem 22.8.8 that

$$\left\| \left((\lambda I - A)^{-1} \right)^n \right\| \leq \frac{M}{\lambda^n}$$

From Lemma 22.8.16 and the fact the adjoint of a bounded linear operator preserves the norm,

$$\frac{M}{\lambda^n} \geq \left\| \left(\left((\lambda I - A)^{-1} \right)^n \right)^* \right\| = \left\| \left(\left((\lambda I - A)^{-1} \right)^* \right)^n \right\| = \left\| \left((\lambda I - A^*)^{-1} \right)^n \right\|$$

and so by Theorem 22.8.8 again it follows A^* generates a continuous semigroup, $T(t)$ which satisfies $||T(t)|| \leq M$. I need to identify $T(t)$ with $S^*(t)$. However, from the proof of Theorem 22.8.8 and Lemma 22.8.16, it follows that for $x^* \in D(A^*)$ and a suitable sequence $\{\lambda_n\}$,

$$T(t)x^*(y) = \lim_{n\to\infty} e^{-\lambda_n t} \sum_{k=0}^{\infty} \frac{t^k \left(\lambda_n^2 (\lambda_n I - A^*)^{-1} \right)^k}{k!} x^*(y)$$

$$= \lim_{n\to\infty} e^{-\lambda_n t} \sum_{k=0}^{\infty} \frac{t^k \left(\left(\lambda_n^2 (\lambda_n I - A)^{-1} \right)^k \right)^*}{k!} x^*(y)$$

$$= \lim_{n\to\infty} x^* \left(e^{-\lambda_n t} \left(\sum_{k=0}^{\infty} \frac{t^k \left(\left(\lambda_n^2 (\lambda_n I - A)^{-1} \right)^k \right)}{k!} y \right) \right) = x^*(S(t)y) = S^*(t)x^*(y).$$

Therefore, since y is arbitrary, $S^*(t) = T(t)$ on $x \in D(A^*)$ a dense set and this shows the two are equal. In particular, $S^*(t)$ is a semigroup because $T(t)$ is. This proves the proposition in the case where $S(t)$ is also bounded.

Next only assume $S(t)$ is a continuous semigroup. Then by Proposition 22.8.5 there exists $\alpha > 0$ such that $\|S(t)\| \leq Me^{\alpha t}$. Then consider the operator $-\alpha I + A$ and the bounded semigroup $e^{-\alpha t} S(t)$. For $x \in D(A)$

$$\lim_{h\to 0+} \frac{e^{-\alpha h} S(h)x - x}{h} = \lim_{h\to 0+} \left(e^{-\alpha h} \frac{S(h)x - x}{h} + \frac{e^{-\alpha h} - 1}{h} x \right) = -\alpha x + Ax$$

Thus $-\alpha I + A$ generates $e^{-\alpha t} S(t)$ and it follows from the first part that $-\alpha I + A^*$ generates the semigroup $e^{-\alpha t} S^*(t)$. Thus

$$\begin{aligned} -\alpha x + A^* x &= \lim_{h\to 0+} \frac{e^{-\alpha h} S^*(h)x - x}{h} \\ &= \lim_{h\to 0+} \left(e^{-\alpha h} \frac{S^*(h)x - x}{h} + \frac{e^{-\alpha h} - 1}{h} x \right) = -\alpha x + \lim_{h\to 0+} \frac{S^*(h)x - x}{h} \end{aligned}$$

showing that A^* generates $S^*(t)$. It follows from Proposition 22.8.5 that A^* is closed and densely defined. It is obvious $S^*(t)$ is a semigroup. Why is it continuous? This also follows from the first part of the argument which establishes that $t \to e^{-\alpha t} S^*(t)x$ is continuous. ■

22.9 Exercises

1. For $f,g \in C([0,1])$ let $(f,g) = \int_0^1 f(x)\overline{g(x)}dx$. Is this an inner product space? Is it a Hilbert space? What does the Cauchy Schwarz inequality say in this context?

2. Let S denote the unit sphere in a Banach space X, $S \equiv \{x \in X : \|x\| = 1\}$. Show that if Y is a Banach space, then $A \in \mathscr{L}(X,Y)$ is compact if and only if $A(S)$ is precompact, $\overline{A(S)}$ is compact. $A \in \mathscr{L}(X,Y)$ is said to be compact if whenever B is a bounded subset of X, it follows $\overline{A(B)}$ is a compact subset of Y. In words, A takes bounded sets to precompact sets.

3. ↑ Show that $A \in \mathscr{L}(X,Y)$ is compact if and only if A^* is compact. **Hint:** Use the result of 2 and the Ascoli Arzela theorem to argue that for S^* the unit ball in X', there is a subsequence, $\{y_n^*\} \subseteq S^*$ such that y_n^* converges uniformly on the compact set, $\overline{A(S)}$. Thus $\{A^* y_n^*\}$ is a Cauchy sequence in X'. To get the other implication, apply the result just obtained for the operators A^* and A^{**}. Then use results about the embedding of a Banach space into its double dual space.

4. Prove the parallelogram identity, $|x+y|^2 + |x-y|^2 = 2|x|^2 + 2|y|^2$. Next suppose $(X, \|\ \|)$ is a real normed linear space and the parallelogram identity holds. Can it be concluded there exists an inner product $(\cdot,\cdot)$ such that $\|x\| = (x,x)^{1/2}$?

5. Let K be a closed, bounded and convex set in $\mathbb{R}^n$ and let $\boldsymbol{f} : K \to \mathbb{R}^n$ be continuous and let $\boldsymbol{y} \in \mathbb{R}^n$. Show using the Brouwer fixed point theorem there exists a point $\boldsymbol{x} \in K$ such that $P(\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{x}) = \boldsymbol{x}$. Next show that $(\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}), \boldsymbol{z} - \boldsymbol{x}) \leq 0$ for all $\boldsymbol{z} \in K$. The existence of this $\boldsymbol{x}$ is known as Browder's lemma and it has great significance in the study of certain types of nolinear operators. Now suppose $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^n$ is continuous and satisfies $\lim_{|\boldsymbol{x}|\to\infty} \frac{(\boldsymbol{f}(\boldsymbol{x}),\boldsymbol{x})}{|\boldsymbol{x}|} = \infty$.Show using Browder's lemma that $\boldsymbol{f}$ is onto.

6. Show that every inner product space is uniformly convex. This means that if x_n, y_n are vectors whose norms are no larger than 1 and if $\|x_n + y_n\| \to 2$, then $\|x_n - y_n\| \to 0$. More precisely, for every $\varepsilon > 0$, there is a $\delta > 0$ such that if $\|x+y\| > 2-\delta$ for $\|x\|, \|y\|$ both 1, then $\|x-y\| < \varepsilon$.

7. Let H be separable and let S be an orthonormal set. Show S is countable. **Hint:** How far apart are two elements of the orthonormal set?

8. Suppose $\{x_1, \cdots, x_m\}$ is a linearly independent set of vectors in a normed linear space. Show $\text{span}(x_1, \cdots, x_m)$ is a closed subspace. Also show every orthonormal set of vectors is linearly independent.

9. Show every Hilbert space, separable or not, has a maximal orthonormal set of vectors.

10. ↑ Prove Bessel's inequality, which says that if $\{x_n\}_{n=1}^{\infty}$ is an orthonormal set in H, then for all $x \in H$, $\|x\|^2 \geq \sum_{k=1}^{\infty} |(x,x_k)|^2$. **Hint:** Show that if $M = \text{span}(x_1, \cdots, x_n)$, then $Px = \sum_{k=1}^{n} x_k (x, x_k)$. Then observe $\|x\|^2 = \|x - Px\|^2 + \|Px\|^2$.

11. ↑ Show S is a maximal orthonormal set if and only if $\text{span}(S)$ is dense in H, where $\text{span}(S)$ is defined as $\text{span}(S) \equiv \{\text{all finite linear combinations of elements of } S\}$.

12. ↑ Suppose $\{x_n\}_{n=1}^{\infty}$ is a maximal orthonormal set. Show that $x = \sum_{n=1}^{\infty}(x,x_n)x_n \equiv \lim_{N\to\infty}\sum_{n=1}^{N}(x,x_n)x_n$ and $\|x\|^2 = \sum_{i=1}^{\infty}|(x,x_i)|^2$. Show $(x,y) = \sum_{n=1}^{\infty}(x,x_n)(\overline{y,x_n})$. **Hint:** For the last part of this, you might proceed as follows. Show $((x,y)) \equiv \sum_{n=1}^{\infty}(x,x_n)(\overline{y,x_n})$ is a well defined inner product on the Hilbert space which delivers the same norm as the original inner product. Then you could verify that there exists a formula for the inner product in terms of the norm and conclude the two inner products, $(\cdot,\cdot)$ and $((\cdot,\cdot))$ must coincide.

13. Suppose X is an infinite dimensional Banach space and suppose $\{x_1 \cdots x_n\}$ are linearly independent with $\|x_i\| = 1$. By Problem 8 span $(x_1 \cdots x_n) \equiv X_n$ is a closed linear subspace of X. Now let $z \notin X_n$ and pick $y \in X_n$ such that $\|z-y\| \leq 2\,\mathrm{dist}(z,X_n)$ and let $x_{n+1} = \frac{z-y}{\|z-y\|}$. Show the sequence $\{x_k\}$ satisfies $\|x_n - x_k\| \geq 1/2$ whenever $k < n$. Now show the unit ball $\{x \in X : \|x\| \leq 1\}$ in a normed linear space is compact if and only if X is finite dimensional. **Hint:** $\left\| \frac{z-y}{\|z-y\|} - x_k \right\| = \left\| \frac{z-y-x_k\|z-y\|}{\|z-y\|} \right\|$.

14. Show that if A is a self adjoint operator on a Hilbert space and $Ay = \lambda y$ for λ a complex number and $y \neq 0$, then λ must be real. Also verify that if A is self adjoint and $Ax = \mu x$ while $Ay = \lambda y$, then if $\mu \neq \lambda$, it must be the case that $(x,y) = 0$.

15. Theorem 22.8.11 gives the the existence and uniqueness for an evolution equation of the form $y' - \Lambda y = g,\ y(0) = y_0 \in D(\Lambda)$ where g is in $C^1(0,\infty;H)$ for H a Banach space. Recall Λ was the generator of a continuous semigroup $S(h)$. Generalize this to an equation of the form

$$y' - \Lambda y = g + Ly,\ y(0) = y_0 \in H$$

where L is a continuous linear map. **Hint:** You might consider $\Lambda + L$ and show it generates a continuous semigroup. Then apply the theorem.

16. Generalize Theorem 22.8.11 in case you know that for each $t > 0, S(t)x \in D(\Lambda)$. You might see about removing the differentiability of g as a requirement and maybe the assumption that $y_0 \in D(\Lambda)$. Analytic semigroups have this property. There we typically start with the closed operator and construct the semigroup $S(t)$ using methods from complex analysis.

Chapter 23

Representation Theorems

23.1 Radon Nikodym Theorem

This chapter is on various representation theorems. The first theorem, the Radon Nikodym Theorem, is a representation theorem for one measure in terms of another. This important theorem represents one measure in terms of another. It is Theorem 10.13.7 on Page 309. For a very different approach, see [49] which has a nice proof due to Von Neumann which is based not on the Hahn decomposition of a signed measure, but on the Riesz representation theorem in Hilbert space.

Definition 23.1.1 *Let μ and λ be two measures defined on a σ-algebra $\mathcal{S}$, of subsets of a set, Ω. λ is absolutely continuous with respect to μ,written as $\lambda \ll \mu$, if $\lambda(E) = 0$ whenever $\mu(E) = 0$. A complex measure λ defined on a σ-algebra $\mathcal{S}$ is one which has the property that if the E_i are distinct and measurable, then $\lambda(\cup_i E_i) = \sum_i \lambda(E_i)$. It is a complex measure because each $\lambda(E_i) \in \mathbb{C}$.*

Recall Corollary10.13.11 on Page 309 which involves the case where λ is a signed measure. I am stating it next for convenience.

Corollary 23.1.2 *Let μ be a finite measure and λ a signed measure ($\lambda(E) \in \mathbb{R}$) with $\lambda \ll \mu$ meaning that if $\mu(E) = 0$ then $\lambda(E) = 0$. Then there exists $h \in L^1$ such that $\lambda(E) = \int_E h d\mu$.*

There is an easy corollary to this which includes complex measures.

Theorem 23.1.3 *Let λ be a complex measure and $\lambda \ll \mu$ for μ a finite measure. Then there exists $h \in L^1$ such that $\lambda(E) = \int_E h d\mu$.*

Proof: Let $(\mathrm{Re}\,\lambda)(E) = \mathrm{Re}(\lambda(E))$ with $\mathrm{Im}\,\lambda$ defined similarly. Then these are signed measures and so there are functions f_1, f_2 in L^1 such that $\mathrm{Re}\,\lambda(E) = \int_E f_1 d\mu, \mathrm{Im}\,\lambda(E) = \int_E f_2 d\mu$. Then $h \equiv f_1 + i f_2$ satisfies the necessary condition. ■

More general versions of the Radon Nikodym theorem available. To see one of these, one can read the treatment in Hewitt and Stromberg [26]. This involves the notion of decomposable measure spaces, a generalization of σ finite.

23.2 Vector Measures

A vector measure is a generalization of a signed or complex measure and it can have values in any topological vector space. Whole books have been written on this subject. See for example the book by Diestal and Uhl [13] titled Vector Measures. I will emphasize only normed linear spaces. This section is about representing one of these measures with respect to its total variation.

Definition 23.2.1 *Let $(V, \|\cdot\|)$ be a normed linear space and let $(\Omega, \mathcal{S})$ be a measure space. A function $\mu : \mathcal{S} \to V$ is a vector measure if μ is countably additive. That is, if $\{E_i\}_{i=1}^{\infty}$ is a sequence of disjoint sets of $\mathcal{S}$,*

$$\mu(\cup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} \mu(E_i).$$

A signed measure is an example as in Definition 10.13.2 on Page 307.

Note that it makes sense to take finite sums because it is given that μ has values in a vector space in which vectors can be summed. In the above, $\mu(E_i)$ is a vector. It might be a point in $\mathbb{R}^p$ or in any other vector space. In many of the most important applications, it is a vector in some sort of function space which may be infinite dimensional. The infinite sum has the usual meaning. That is $\sum_{i=1}^{\infty}\mu(E_i) = \lim_{n\to\infty}\sum_{i=1}^{n}\mu(E_i)$ where the limit takes place relative to the norm on V.

Definition 23.2.2 *Let $(\Omega,\mathcal{S})$ be a measure space and let μ be a vector measure defined on $\mathcal{S}$. A subset, $\pi(E)$, of $\mathcal{S}$ is called a partition of E if $\pi(E)$ consists of finitely many disjoint sets of $\mathcal{S}$ and $\cup\pi(E) = E$. Let*

$$|\mu|(E) = \sup\{\sum_{F\in\pi(E)}\|\mu(F)\| : \pi(E) \text{ is a partition of } E\}.$$

$|\mu|$ is called the total variation of μ.

The next theorem may seem a little surprising. It states that, if finite, the total variation is a nonnegative measure.

Theorem 23.2.3 *If $|\mu|(\Omega) < \infty$, then $|\mu|$ is a measure on $\mathcal{S}$. Even if $|\mu|(\Omega) = \infty$, $|\mu|(\cup_{i=1}^{\infty}E_i) \le \sum_{i=1}^{\infty}|\mu|(E_i)$. That is $|\mu|$ is always subadditive and $|\mu|(A) \le |\mu|(B)$ whenever $A,B\in\mathcal{S}$ with $A\subseteq B$. In earlier terminology, $|\mu|$ is an outer measure.*

Proof: Consider the last claim. Let $a < |\mu|(A)$ and let $\pi(A)$ be a partition of A such that $a < \sum_{F\in\pi(A)}\|\mu(F)\|$. Then $\pi(A)\cup\{B\setminus A\}$ is a partition of B and

$$|\mu|(B) \ge \sum_{F\in\pi(A)}\|\mu(F)\| + \|\mu(B\setminus A)\| > a.$$

Since this is true for all such a, it follows $|\mu|(B) \ge |\mu|(A)$ as claimed.

Let $\{E_j\}_{j=1}^{\infty}$ be a sequence of disjoint sets of $\mathcal{S}$ and let $E_\infty = \cup_{j=1}^{\infty}E_j$. Then letting $a < |\mu|(E_\infty)$, it follows from the definition of total variation there exists a partition of E_∞, $\pi(E_\infty) = \{A_1,\cdots,A_n\}$ such that $a < \sum_{i=1}^{n}\|\mu(A_i)\|$. Also, $A_i = \cup_{j=1}^{\infty}A_i\cap E_j$ and so by the triangle inequality, $\|\mu(A_i)\| \le \sum_{j=1}^{\infty}\|\mu(A_i\cap E_j)\|$. Therefore, by the above, and either Fubini's theorem or Lemma 2.5.4 on Page 65,

$$a < \sum_{i=1}^{n}\overbrace{\sum_{j=1}^{\infty}\|\mu(A_i\cap E_j)\|}^{\ge\|\mu(A_i)\|} = \sum_{j=1}^{\infty}\sum_{i=1}^{n}\|\mu(A_i\cap E_j)\| \le \sum_{j=1}^{\infty}|\mu|(E_j)$$

because $\{A_i\cap E_j\}_{i=1}^{n}$ is a partition of E_j.

Since a is arbitrary, this shows $|\mu|(\cup_{j=1}^{\infty}E_j) \le \sum_{j=1}^{\infty}|\mu|(E_j)$. If the sets, E_j are not disjoint, let $F_1 = E_1$ and if F_n has been chosen, let $F_{n+1} \equiv E_{n+1}\setminus\cup_{i=1}^{n}E_i$. Thus the sets, F_i are disjoint and $\cup_{i=1}^{\infty}F_i = \cup_{i=1}^{\infty}E_i$. Therefore,

$$|\mu|(\cup_{j=1}^{\infty}E_j) = |\mu|(\cup_{j=1}^{\infty}F_j) \le \sum_{j=1}^{\infty}|\mu|(F_j) \le \sum_{j=1}^{\infty}|\mu|(E_j)$$

and proves $|\mu|$ is always subadditive as claimed, regardless of whether $|\mu|(\Omega) < \infty$.

Now suppose $|\mu|(\Omega) < \infty$ and let E_1 and E_2 be sets of $\mathscr{S}$ such that $E_1 \cap E_2 = \emptyset$ and let $\{A_1^i \cdots A_{n_i}^i\} = \pi(E_i)$, a partition of E_i which is chosen such that

$$|\mu|(E_i) - \varepsilon < \sum_{j=1}^{n_i} \left\|\mu(A_j^i)\right\| \quad i = 1,2.$$

Such a partition exists because of the definition of the total variation. Considering the sets which are contained in either of $\pi(E_1)$ or $\pi(E_2)$, it follows this finite collection of sets is a partition of $E_1 \cup E_2$ denoted by $\pi(E_1 \cup E_2)$. Then by the above inequality and the definition of total variation,

$$|\mu|(E_1 \cup E_2) \geq \sum_{F \in \pi(E_1 \cup E_2)} \|\mu(F)\| > |\mu|(E_1) + |\mu|(E_2) - 2\varepsilon,$$

which shows that since $\varepsilon > 0$ was arbitrary, $|\mu|(E_1 \cup E_2) \geq |\mu|(E_1) + |\mu|(E_2)$. By induction, whenever the E_i are disjoint, $|\mu|(\cup_{j=1}^n E_j) \geq \sum_{j=1}^n |\mu|(E_j)$. Therefore,

$$\sum_{j=1}^{\infty} |\mu|(E_j) \geq |\mu|(\cup_{j=1}^{\infty} E_j) \geq |\mu|(\cup_{j=1}^n E_j) \geq \sum_{j=1}^n |\mu|(E_j).$$

Now let $n \to \infty$. Thus, $|\mu|(\cup_{j=1}^{\infty} E_j) = \sum_{j=1}^{\infty} |\mu|(E_j)$ which shows that $|\mu|$ is a measure as claimed. ■

The following corollary is interesting. It concerns the case that μ is only finitely additive.

Corollary 23.2.4 *Suppose $(\Omega, \mathscr{F})$ is a set with a σ algebra of subsets $\mathscr{F}$ and suppose $\mu : \mathscr{F} \to \mathbb{C}$ is only finitely additive. That is, $\mu\left(\cup_{i=1}^n E_i\right) = \sum_{i=1}^n \mu(E_i)$ whenever the E_i are disjoint. Then $|\mu|$, defined in the same way as above, is also finitely additive provided $|\mu|$ is finite.*

Proof: Say $E \cap F = \emptyset$ for $E, F \in \mathscr{F}$. Let $\pi(E), \pi(F)$ suitable partitions for which the following holds.

$$|\mu|(E \cup F) \geq \sum_{A \in \pi(E)} |\mu(A)| + \sum_{B \in \pi(F)} |\mu(B)| \geq |\mu|(E) + |\mu|(F) - 2\varepsilon.$$

Since ε is arbitrary, $|\mu|(E \cap F) \geq |\mu|(E) + |\mu|(F)$. Similar considerations apply to any finite union of disjoint sets. That is, if the E_i are disjoint, then $|\mu|\left(\cup_{i=1}^n E_i\right) \geq \sum_{i=1}^n |\mu|(E_i)$.

Now let $E = \cup_{i=1}^n E_i$ where the E_i are disjoint. Then letting $\pi(E)$ be a suitable partition of E,

$$|\mu|(E) - \varepsilon \leq \sum_{F \in \pi(E)} |\mu(F)|,$$

it follows that

$$|\mu|(E) \leq \varepsilon + \sum_{F \in \pi(E)} |\mu(F)| = \varepsilon + \sum_{F \in \pi(E)} \left|\sum_{i=1}^n \mu(F \cap E_i)\right|$$

$$\leq \varepsilon + \sum_{i=1}^n \sum_{F \in \pi(E)} |\mu(F \cap E_i)| \leq \varepsilon + \sum_{i=1}^n |\mu|(E_i)$$

Since ε is arbitrary, this shows $|\mu|\left(\cup_{i=1}^{n}E_i\right) \leq \sum_{i=1}^{n}|\mu|(E_i)$. Thus $|\mu|$ is finitely additive. ■

In the case that λ is a complex measure, it is always the case that $|\lambda|(\Omega) < \infty$. First is a lemma.

Lemma 23.2.5 *Suppose λ is a real valued measure called a signed measure (Definition 10.13.2). Then $|\lambda|$ is a finite measure.*

Proof: Suppose $\lambda : \mathscr{F} \to \mathbb{R}$ is a vector measure (signed measure by Definition 10.13.2). By the Hahn decomposition, Theorem 10.13.5 on Page 307, $\Omega = P \cup N$ where P is a positive set and N is a negative one. Then on N, $-\lambda$ acts like a measure in the sense that if $A \subseteq B$ and A, B measurable subsets of N, then $-\lambda(A) \leq -\lambda(B)$. Similarly λ is a measure on P.

$$\sum_{F\in\pi(\Omega)}|\lambda(F)| \leq \sum_{F\in\pi(\Omega)}(|\lambda(F\cap P)|+|\lambda(F\cap N)|)$$

$$= \sum_{F\in\pi(\Omega)}\lambda(F\cap P)+\sum_{F\in\pi(\Omega)}-\lambda(F\cap N)$$

$$= \lambda\left(\left(\cup_{F\in\pi(\Omega)}F\right)\cap P\right)+-\lambda\left(\left(\cup_{F\in\pi(\Omega)}F\right)\cap N\right) \leq \lambda(P)+|\lambda(N)|$$

It follows that $|\lambda|(\Omega) < \lambda(P)+|\lambda(N)|$ and so $|\lambda|$ has finite total variation. ■

Theorem 23.2.6 *Suppose λ is a complex measure on $(\Omega, \mathscr{S})$ where $\mathscr{S}$ is a σ algebra of subsets of Ω. Then $|\lambda|(\Omega) < \infty$.*

Proof: If λ is a vector measure with values in $\mathbb{C}$, $\operatorname{Re}\lambda$ and $\operatorname{Im}\lambda$ have values in $\mathbb{R}$. Then

$$\begin{aligned}\sum_{F\in\pi(\Omega)}|\lambda(F)| &\leq \sum_{F\in\pi(\Omega)}|\operatorname{Re}\lambda(F)|+|\operatorname{Im}\lambda(F)|\\ &= \sum_{F\in\pi(\Omega)}|\operatorname{Re}\lambda(F)|+\sum_{F\in\pi(\Omega)}|\operatorname{Im}\lambda(F)|\\ &\leq |\operatorname{Re}\lambda|(\Omega)+|\operatorname{Im}\lambda|(\Omega) < \infty\end{aligned}$$

thanks to Lemma 23.2.5. ■

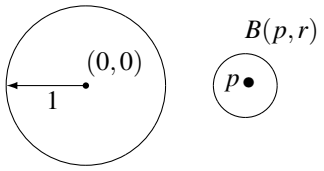
The following corollary is about representing a complex measure λ in terms of its total variation $|\lambda|$. It is like representing a complex number in the form $re^{i\theta}$. In particular, I want to show that in the Radon Nikodym theorem $\left|\frac{d\lambda}{d\mu}\right| = 1$ a.e. First is a lemma which is interesting for its own sake and shows $\left|\frac{d\lambda}{d\mu}\right| \leq 1$.

Lemma 23.2.7 *Suppose $(\Omega, \mathscr{S}, \mu)$ is a measure space and f is a function in $L^1(\Omega, \mu)$ with the property that*

$$\left|\int_E f d\mu\right| \leq \mu(E)$$

for all $E \in \mathscr{S}$. Then $|f| \leq 1$ a.e.

Proof of the lemma: Consider the following picture where $B(p,r)\cap B(0,1) = \emptyset$.



Let $E = f^{-1}(B(p,r))$. In fact $\mu(E) = 0$. If $\mu(E) \neq 0$ then

$$\left|\frac{1}{\mu(E)}\int_E f\,d\mu - p\right| = \left|\frac{1}{\mu(E)}\int_E (f-p)d\mu\right| \leq \frac{1}{\mu(E)}\int_E |f-p|d\mu < r$$

because on E, $|f(\omega)-p| < r$. Hence $\frac{1}{\mu(E)}\int_E f\,d\mu$ is closer to p than r and so

$$\left|\frac{1}{\mu(E)}\int_E f d\mu\right| > 1.$$

Refer to the picture. However, this contradicts the assumption of the lemma. It follows $\mu(E) = 0$. Since the set of complex numbers z such that $|z| > 1$ is an open set, it equals the union of countably many balls, $\{B_i\}_{i=1}^{\infty}$. Therefore,

$$\mu\left(f^{-1}(\{z \in \mathbb{C} : |z| > 1\}\right) = \mu\left(\cup_{k=1}^{\infty} f^{-1}(B_k)\right) \leq \sum_{k=1}^{\infty} \mu\left(f^{-1}(B_k)\right) = 0.$$

Thus $|f(\omega)| \leq 1$ a.e. as claimed. ■

Note that the above argument would work with essentially no change if $\mathbb{C}$ were replaced with V a separable normed linear space.

Corollary 23.2.8 *Let λ be a complex vector measure with $|\lambda|(\Omega) < \infty$.[1] Then there exists a unique $f \in L^1(\Omega)$ such that $\lambda(E) = \int_E f d|\lambda|$. Furthermore, $|f| = 1$ for $|\lambda|$ a.e. This is called the polar decomposition of λ. We write $d\lambda = f\,d|\lambda|$ sometimes.*

Proof: Letting $\mu = |\lambda|$ in Theorem 23.1.3, the first claim follows because $\lambda \ll |\lambda|$ and so such an L^1 function exists and is unique. It is required to show $|f| = 1$ a.e. If $|\lambda|(E) \neq 0$, $\left|\frac{\lambda(E)}{|\lambda|(E)}\right| = \left|\frac{1}{|\lambda|(E)}\int_E f\,d|\lambda|\right| \leq 1$. Therefore by Lemma 23.2.7, $|f| \leq 1$, $|\lambda|$ a.e. Now let

$$E_n = \left[|f| \leq 1 - \frac{1}{n}\right].$$

Let $\{F_1, \cdots, F_m\}$ be a partition of E_n. Then

$$\sum_{i=1}^{m} |\lambda(F_i)| = \sum_{i=1}^{m} \left|\int_{F_i} f d|\lambda|\right| \leq \sum_{i=1}^{m} \int_{F_i} |f|\,d|\lambda|$$

$$\leq \sum_{i=1}^{m} \int_{F_i} \left(1 - \frac{1}{n}\right) d|\lambda| = \sum_{i=1}^{m} \left(1 - \frac{1}{n}\right)|\lambda|(F_i) = |\lambda|(E_n)\left(1 - \frac{1}{n}\right).$$

Then taking the supremum over all partitions, $|\lambda|(E_n) \leq \left(1 - \frac{1}{n}\right)|\lambda|(E_n)$ which shows $|\lambda|(E_n) = 0$. Hence $|\lambda|([|f| < 1]) = 0$ because $[|f| < 1] = \cup_{n=1}^{\infty} E_n$. ■

Next is a specific case which leads to complex measures.

[1] As proved above, the assumption that $|\lambda|(\Omega) < \infty$ is redundant.

Corollary 23.2.9 *Suppose $(\Omega, \mathcal{S})$ is a measure space and μ is a finite nonnegative measure on $\mathcal{S}$. Then for $h \in L^1(\mu)$, define a complex measure, λ by $\lambda(E) \equiv \int_E h d\mu$. Then $|\lambda|(E) = \int_E |h| d\mu$. Furthermore, $|h| = \overline{g}h$ where $g d|\lambda|$ is the polar decomposition of λ, defined by $\lambda(E) = \int_E g d|\lambda|$.*

Proof: From Corollary 23.2.8, there exists g such that $|g| = 1, |\lambda|$ a.e. and for all $E \in \mathcal{S}$

$$\lambda(E) = \int_E g d|\lambda|, \ \lambda(E) \equiv \int_E h d\mu, \text{ so } \int_E h d\mu = \int_E g d|\lambda| \tag{23.1}$$

Since $|g| = 1$, there is a sequence s_n of simple functions converging pointwise to $\overline{g}$ with $|s_n| \leq 1$. (Approximate the positive and negative parts of the real and imaginary parts of $\overline{g}$ with an increasing sequence of simple functions. Then assemble these to get s_n. See Theorem 10.7.6 on Page 291.) Then from 23.1, $\int_E g s_n d|\lambda| = \int_E s_n h d\mu$. Passing to the limit using the dominated convergence theorem, $\int_E d|\lambda| = \int_E \overline{g} h d\mu$. It follows $\overline{g}h \geq 0$ a.e. and $|\overline{g}| = 1$ *a.e.* Therefore, $|h| = |\overline{g}h| = \overline{g}h$. It follows from the above, that

$$|\lambda|(E) = \int_E d|\lambda| = \int_E \overline{g} h d\mu = \int_E |h| d\mu \ \blacksquare$$

Formally: If $d\lambda = h d\mu$, and $d\lambda = g d|\lambda|, |g| = 1$, then $d|\lambda| = |h| d\mu$ and so you might expect to have $d\lambda = g|h| d\mu$ so $g|h| = h$ and so $|h| = \overline{g}h$. That which should be true is. Emphasizing the most significant part of this, if $d\lambda = h d\mu$, then $d|\lambda| = |h| d\mu$.

23.3 Representation for the Dual Space of L^p

Recall the concept of the dual space of a Banach space in the chapter on Banach space starting on Page 557. The next topic deals with the dual space of L^p for $p \geq 1$ in the case where the measure space is σ finite or finite. In what follows $q = \infty$ if $p = 1$ and otherwise, $\frac{1}{p} + \frac{1}{q} = 1$. In what follows, $|\cdot|$ is the usual norm on $\mathbb{C}$.

Theorem 23.3.1 *(Riesz representation theorem) Let $\infty > p > 1$ and let $(\Omega, \mathcal{S}, \mu)$ be a finite measure space. If $\Lambda \in (L^p(\Omega))'$, then there exists a unique $h \in L^q(\Omega)$ such that*

$$\Lambda f = \int_\Omega h f d\mu.$$

This function satisfies $\|h\|_q = \|\Lambda\|$ where $\|\Lambda\|$ is the operator norm of Λ.

Proof: (Uniqueness) If h_1 and h_2 both represent Λ, consider

$$f = |h_1 - h_2|^{q-2}(\overline{h_1} - \overline{h_2}),$$

where $\overline{h}$ denotes complex conjugation. By Holder's inequality, it is easy to see that $f \in L^p(\Omega)$. Thus $0 = \Lambda f - \Lambda f = \int h_1 |h_1 - h_2|^{q-2}(\overline{h_1} - \overline{h_2}) - h_2 |h_1 - h_2|^{q-2}(\overline{h_1} - \overline{h_2}) d\mu = \int |h_1 - h_2|^q d\mu$. Therefore $h_1 = h_2$ and this proves uniqueness in every case regarless whether μ is finite.

Now let $\lambda(E) = \Lambda(\mathcal{X}_E)$. Since this is a finite measure space, $\mathcal{X}_E$ is an element of $L^p(\Omega)$ and so it makes sense to write $\Lambda(\mathcal{X}_E)$. In fact λ is a complex measure having finite total variation. First I show that it is a measure. Then by Theorem 23.2.6, λ has finite total variation.

If $\{E_i\}_{i=1}^{\infty}$ is a sequence of disjoint sets of $\mathscr{S}$, let $F_n = \cup_{i=1}^{n} E_i$, $F = \cup_{i=1}^{\infty} E_i$. Then by the Dominated Convergence theorem, $\|\mathscr{X}_{F_n} - \mathscr{X}_F\|_p \to 0$. Therefore, by continuity of Λ,

$$\lambda(F) \equiv \Lambda(\mathscr{X}_F) = \lim_{n\to\infty} \Lambda(\mathscr{X}_{F_n}) = \lim_{n\to\infty} \sum_{k=1}^{n} \Lambda(\mathscr{X}_{E_k}) = \sum_{k=1}^{\infty} \lambda(E_k).$$

This shows λ is a complex measure. Since a similar theorem will be proved in which λ has values in an infinite dimensional space, I will prove this directly without using Theorem 23.2.6. Let $A_1, \cdots, A_n$ be a partition of Ω. $|\Lambda\mathscr{X}_{A_i}| = w_i(\Lambda\mathscr{X}_{A_i}) = \Lambda(w_i\mathscr{X}_{A_i})$ for some $w_i \in \mathbb{C}$, $|w_i| = 1$. Thus

$$\begin{aligned} \sum_{i=1}^{n} |\lambda(A_i)| &= \sum_{i=1}^{n} |\Lambda(\mathscr{X}_{A_i})| = \Lambda(\sum_{i=1}^{n} w_i \mathscr{X}_{A_i}) \\ &\leq \|\Lambda\|(\int |\sum_{i=1}^{n} w_i \mathscr{X}_{A_i}|^p d\mu)^{\frac{1}{p}} = \|\Lambda\|(\int_{\Omega} d\mu)^{\frac{1}{p}} = \|\Lambda\|\mu(\Omega)^{\frac{1}{p}}. \end{aligned}$$

This is because if $x \in \Omega$, x is contained in exactly one of the A_i and so the absolute value of the sum in the first integral above is equal to 1. Therefore $|\lambda|(\Omega) < \infty$ because this was an arbitrary partition. with $|\lambda|$ finite.

It is also clear from the definition of λ that $\lambda \ll \mu$. Therefore, by the Radon Nikodym theorem, there exists $h \in L^1(\Omega)$ with $\lambda(E) = \int_E h d\mu = \Lambda(\mathscr{X}_E)$. Actually $h \in L^q$ and satisfies the other conditions above. This is shown next.

Let $s = \sum_{i=1}^{m} c_i \mathscr{X}_{E_i}$ be a simple function. Then since Λ is linear,

$$\Lambda(s) = \sum_{i=1}^{m} c_i \Lambda(\mathscr{X}_{E_i}) = \sum_{i=1}^{m} c_i \int_{E_i} h d\mu = \int h s d\mu. \tag{23.2}$$

Claim: If f is uniformly bounded and measurable, then

$$\Lambda(f) = \int h f d\mu.$$

Proof of claim: Since f is bounded and measurable, there exists a sequence of simple functions, $\{s_n\}$ which converges to f pointwise and in $L^p(\Omega)$, $|s_n| \leq |f|$. This follows from Theorem 9.1.6 on Page 244 upon breaking f up into positive and negative parts of real and complex parts. In fact this theorem gives uniform convergence. Then

$$\Lambda(f) = \lim_{n\to\infty} \Lambda(s_n) = \lim_{n\to\infty} \int h s_n d\mu = \int h f d\mu,$$

the first equality holding because of continuity of Λ, the second following from 23.2 and the third holding by the dominated convergence theorem.

This is a very nice formula but it still has not been shown that $h \in L^q(\Omega)$.

Let $E_n = \{x : |h(x)| \leq n\}$. Thus $|h\mathscr{X}_{E_n}| \leq n$. Then

$$|h\mathscr{X}_{E_n}|^{q-2}(\overline{h}\mathscr{X}_{E_n}) \in L^p(\Omega).$$

By the claim, it follows that

$$\|h\mathscr{X}_{E_n}\|_q^q = \int h|h\mathscr{X}_{E_n}|^{q-2}(\overline{h}\mathscr{X}_{E_n})d\mu = \Lambda(|h\mathscr{X}_{E_n}|^{q-2}(\overline{h}\mathscr{X}_{E_n}))$$

$$\leq \|\Lambda\| \left\| |h\mathcal{X}_{E_n}|^{q-2}(\overline{h}\mathcal{X}_{E_n}) \right\|_p = \left(\int |h\mathcal{X}_{E_n}|^q d\mu \right)^{1/p} = \|\Lambda\| \, \|h\mathcal{X}_{E_n}\|_q^{\frac{q}{p}},$$

because $q-1 = q/p$ and so it follows that $\|h\mathcal{X}_{E_n}\|_q \leq \|\Lambda\|$. Letting $n \to \infty$, the monotone convergence theorem implies

$$\|h\|_q \leq \|\Lambda\|. \tag{23.3}$$

Now that h has been shown to be in $L^q(\Omega)$, it follows from 23.2 and the density of the simple functions, Theorem 12.2.1 on Page 370, that $\Lambda f = \int hf d\mu$ for all $f \in L^p(\Omega)$. It only remains to verify the last claim that $\|h\|_q = \|\Lambda\|$ not just 23.3. However, from the definition and Holder's inequality and 23.3, $\|\Lambda\| \equiv \sup\{\int hf : \|f\|_p \leq 1\} \leq \|h\|_q \leq \|\Lambda\|$ ■

To represent elements of the dual space of $L^1(\Omega)$, another Banach space is needed.

Definition 23.3.2

Let $(\Omega, \mathcal{S}, \mu)$ be a measure space. $L^\infty(\Omega)$ is the vector space of measurable functions such that for some $M > 0$, $|f(x)| \leq M$ for all x outside of some set of measure zero ($|f(x)| \leq M$ a.e.). Define $f = g$ when $f(x) = g(x)$ a.e. and $\|f\|_\infty \equiv \inf\{M : |f(x)| \leq M \text{ a.e.}\}$.

Theorem 23.3.3

$L^\infty(\Omega)$ is a Banach space.

Proof: It is clear that $L^\infty(\Omega)$ is a vector space. Is $\| \|_\infty$ a norm?

Claim: If $f \in L^\infty(\Omega)$, then $|f(x)| \leq ||f||_\infty$ a.e.

Proof of the claim: $\left\{x : |f(x)| \geq ||f||_\infty + n^{-1}\right\} \equiv E_n$ is a set of measure zero according to the definition of $||f||_\infty$. Furthermore, $\{x : |f(x)| > ||f||_\infty\} = \cup_n E_n$ and so it is also a set of measure zero. This verifies the claim.

Now if $||f||_\infty = 0$ it follows that $f(x) = 0$ a.e. Also if

$$f, g \in L^\infty(\Omega)$$

then $|f(x)+g(x)| \leq |f(x)|+|g(x)| \leq ||f||_\infty + ||g||_\infty$ a.e. and so $||f||_\infty + ||g||_\infty$ serves as one of the constants, M in the definition of $||f+g||_\infty$. Therefore, $||f+g||_\infty \leq ||f||_\infty + ||g||_\infty$. Next let c be a number. Then $|cf(x)| = |c|\,|f(x)| \leq |c|\,||f||_\infty$ and so $||cf||_\infty \leq |c|\,||f||_\infty$. Therefore since c is arbitrary, $||f||_\infty = ||c(1/c)f||_\infty \leq \left|\frac{1}{c}\right| ||cf||_\infty$ which implies $|c|\,||f||_\infty \leq ||cf||_\infty$. Thus $\| \|_\infty$ is a norm as claimed.

To verify completeness, let $\{f_n\}$ be a Cauchy sequence in $L^\infty(\Omega)$ and use the above claim to get the existence of a set of measure zero, E_{nm} such that for all $x \notin E_{nm}$,

$$|f_n(x) - f_m(x)| \leq \|f_n - f_m\|_\infty.$$

Let $E = \cup_{n,m} E_{nm}$. Thus $\mu(E) = 0$ and for each $x \notin E$, $\{f_n(x)\}_{n=1}^\infty$ is a Cauchy sequence in $\mathbb{C}$. Let

$$f(x) = \begin{cases} 0 \text{ if } x \in E \\ \lim_{n\to\infty} f_n(x) \text{ if } x \notin E \end{cases} = \lim_{n\to\infty} \mathcal{X}_{E^C}(x) f_n(x).$$

Then f is clearly measurable because it is the limit of measurable functions. If

$$F_n = \{x : |f_n(x)| > \|f_n\|_\infty\}$$

and $F = \cup_{n=1}^\infty F_n$, it follows $\mu(F) = 0$ and that for $x \notin F \cup E$,

$$|f(x)| \leq \liminf_{n\to\infty} |f_n(x)| \leq \liminf_{n\to\infty} \|f_n\|_\infty < \infty$$

because $\{\|f_n\|_\infty\}$ is a Cauchy sequence. ($|\|f_n\|_\infty - \|f_m\|_\infty| \leq \|f_n - f_m\|_\infty$ by the triangle inequality.) Thus $f \in L^\infty(\Omega)$. Let n be large enough that whenever $m > n$, $\|f_m - f_n\|_\infty < \varepsilon$. Then, if $x \notin E$,

$$|f(x) - f_n(x)| = \lim_{m\to\infty} |f_m(x) - f_n(x)| \leq \lim_{m\to\infty} \inf \|f_m - f_n\|_\infty < \varepsilon.$$

Hence $\|f - f_n\|_\infty < \varepsilon$ for all n large enough. ■

The next theorem is the Riesz representation theorem for $\left(L^1(\Omega)\right)'$.

Theorem 23.3.4 *(Riesz representation theorem) Let $(\Omega, \mathscr{S}, \mu)$ be a finite measure space. If $\Lambda \in (L^1(\Omega))'$, then there exists a unique $h \in L^\infty(\Omega)$ such that*

$$\Lambda(f) = \int_\Omega hf \, d\mu$$

for all $f \in L^1(\Omega)$. If h is the function in $L^\infty(\Omega)$ representing $\Lambda \in (L^1(\Omega))'$, then $\|h\|_\infty = \|\Lambda\|$.

Proof: Just as in the proof of Theorem 23.3.1, there exists a unique $h \in L^1(\Omega)$ such that for all simple functions s,

$$\Lambda(s) = \int hs \, d\mu. \tag{23.4}$$

To show $h \in L^\infty(\Omega)$, let $\varepsilon > 0$ be given and let $E = \{x : |h(x)| \geq \|\Lambda\| + \varepsilon\}$. Let $|k| = 1$ and $hk = |h|$. Since the measure space is finite, $k \in L^1(\Omega)$. As in Theorem 23.3.1 let $\{s_n\}$ be a sequence of simple functions converging to k in $L^1(\Omega)$, and pointwise. It follows from the construction in Theorem 9.1.6 on Page 244 that it can be assumed $|s_n| \leq 1$. Therefore

$$\Lambda(k\mathscr{X}_E) = \lim_{n\to\infty} \Lambda(s_n \mathscr{X}_E) = \lim_{n\to\infty} \int_E hs_n d\mu = \int_E hk d\mu$$

where the last equality holds by the Dominated Convergence theorem. Therefore,

$$\begin{aligned} \|\Lambda\|\mu(E) &\geq |\Lambda(k\mathscr{X}_E)| = |\int_\Omega hk\mathscr{X}_E d\mu| = \int_E |h| d\mu \\ &\geq (\|\Lambda\| + \varepsilon)\mu(E). \end{aligned}$$

It follows that $\mu(E) = 0$. Since $\varepsilon > 0$ was arbitrary, $\|\Lambda\| \geq \|h\|_\infty$. Since $h \in L^\infty(\Omega)$, the density of the simple functions in $L^1(\Omega)$ and 23.4 imply

$$\Lambda f = \int_\Omega hf d\mu \ , \ \|\Lambda\| \geq \|h\|_\infty. \tag{23.5}$$

This proves the existence part of the theorem. To verify uniqueness, suppose h_1 and h_2 both represent Λ and let $f \in L^1(\Omega)$ be such that $|f| \leq 1$ and $f(h_1 - h_2) = |h_1 - h_2|$. Then $0 = \Lambda f - \Lambda f = \int (h_1 - h_2) f d\mu = \int |h_1 - h_2| d\mu$. Thus $h_1 = h_2$. Finally, $\|\Lambda\| = \sup\{|\int hf d\mu| : \|f\|_1 \leq 1\} \leq \|h\|_\infty \leq \|\Lambda\|$by 23.5. ■

Next these results are extended to the σ finite case.

Lemma 23.3.5 *Let $(\Omega, \mathscr{S}, \mu)$ be a measure space and suppose there exists a measurable function, r such that $r(x) > 0$ for all x, there exists M such that $|r(x)| < M$ for all x, and $\int r d\mu < \infty$. Then for $\Lambda \in (L^p(\Omega, \mu))'$, $p \geq 1$, there exists $h \in L^q(\Omega, \mu)$, $L^\infty(\Omega, \mu)$ if $p = 1$ such that $\Lambda f = \int hf d\mu$. Also $\|h\| = \|\Lambda\|$. ($\|h\| = \|h\|_q$ if $p > 1$, $\|h\|_\infty$ if $p = 1$). Here $\frac{1}{p} + \frac{1}{q} = 1$.*

Proof: Define a new measure $\widetilde{\mu}$, according to the rule

$$\widetilde{\mu}(E) \equiv \int_E r d\mu. \tag{23.6}$$

Thus $\widetilde{\mu}$ is a finite measure on $\mathscr{S}$. For

$$\Lambda \in (L^p(\mu))', \Lambda(f) = \Lambda\left(r^{1/p}\left(r^{-1/p}f\right)\right) = \tilde{\Lambda}\left(r^{-1/p}f\right)$$

where $\widetilde{\Lambda}(g) \equiv \Lambda\left(r^{1/p}g\right)$. Now $\tilde{\Lambda}$ is in $L^p(\tilde{\mu})'$ because

$$\begin{aligned}\left|\widetilde{\Lambda}(g)\right| &\equiv \left|\Lambda\left(r^{1/p}g\right)\right| \leq \|\Lambda\|\left(\int_\Omega \left|r^{1/p}g\right|^p d\mu\right)^{1/p} \\ &= \|\Lambda\|\left(\int_\Omega |g|^p \overbrace{r d\mu}^{d\tilde{\mu}}\right)^{1/p} = \|\Lambda\|\,\|g\|_{L^p(\widetilde{\mu})}\end{aligned}$$

Therefore, by Theorems 23.3.4 and 23.3.1 there exists a unique $h \in L^q(\widetilde{\mu})$ which represents $\widetilde{\Lambda}$. Here $q = \infty$ if $p = 1$ and satisfies $1/q + 1/p = 1$ otherwise. Then

$$\Lambda(f) = \tilde{\Lambda}\left(r^{-1/p}f\right) = \int_\Omega h f r^{-1/p} r d\mu = \int_\Omega f\left(hr^{1/q}\right) d\mu$$

Now $hr^{1/q} \equiv \tilde{h} \in L^q(\mu)$ since $h \in L^q(\tilde{\mu})$. In case $p = 1, L^q(\widetilde{\mu})$ and $L^q(\mu)$ are exactly the same. In this case you have

$$\Lambda(f) = \tilde{\Lambda}\left(r^{-1}f\right) = \int_\Omega h f r^{-1} r d\mu = \int_\Omega f h d\mu$$

Thus the desired representation holds. Then in any case, $|\Lambda(f)| \leq \left\|\tilde{h}\right\|_{L^q}\|f\|_{L^p}$ so $\|\Lambda\| \leq \left\|\tilde{h}\right\|_{L^q}$. Also, as before,

$$\left\|\tilde{h}\right\|^q_{L^q(\mu)} = \left|\int_\Omega \tilde{h}\left|\tilde{h}\right|^{q-2}\overline{\tilde{h}} d\mu\right| = \left|\Lambda\left(\left|\tilde{h}\right|^{q-2}\overline{\tilde{h}}\right)\right| \leq \|\Lambda\|\left(\int_\Omega \left|\left|\tilde{h}\right|^{q-2}\overline{\tilde{h}}\right|^p d\mu\right)^{1/p}$$

$$= \|\Lambda\|\left(\int_\Omega \left(\left|\tilde{h}\right|^{q/p}\right)^p\right)^{1/p} = \|\Lambda\|\,\|h\|^{q/p}$$

and so $\left\|\tilde{h}\right\|_{L^q(\mu)} \leq \|\Lambda\|$. It works the same for $p = 1$. Thus $\left\|\tilde{h}\right\|_{L^q(\mu)} = \|\Lambda\|$. ■

A situation in which the conditions of the lemma are satisfied is the case where the measure space is σ finite. In fact, you should show this is the only case in which the conditions of the above lemma hold.

Theorem 23.3.6 *(Riesz representation theorem) Let $(\Omega, \mathscr{S}, \mu)$ be σ finite and let $\Lambda \in (L^p(\Omega,\mu))'$, $p \geq 1$. Then there exists a unique $h \in L^q(\Omega,\mu)$, $L^\infty(\Omega,\mu)$ if $p = 1$ such that $\Lambda f = \int h f d\mu$. Also $\|h\| = \|\Lambda\|$. ($\|h\| = \|h\|_q$ if $p > 1$, $\|h\|_\infty$ if $p = 1$). Here $\frac{1}{p} + \frac{1}{q} = 1$.*

Proof: Without loss of generality, assume $\mu(\Omega) = \infty$. By Proposition 10.13.1, either μ is a finite measure or $\mu(\Omega) = \infty$. These are the only two cases. Then let $\{\Omega_n\}$ be a sequence

of disjoint elements of $\mathscr{S}$ having the property that $1 < \mu(\Omega_n) < \infty$, $\cup_{n=1}^{\infty}\Omega_n = \Omega$. Define $r(x) = \sum_{n=1}^{\infty} \frac{1}{n^2} \mathscr{X}_{\Omega_n}(x)\, \mu(\Omega_n)^{-1}$, $\widetilde{\mu}(E) = \int_E r d\mu$. Thus $\int_\Omega r d\mu = \widetilde{\mu}(\Omega) = \sum_{n=1}^{\infty} \frac{1}{n^2} < \infty$ so $\widetilde{\mu}$ is a finite measure. The above lemma gives the existence part of the conclusion of the theorem. Uniqueness is done as before. ■

With the Riesz representation theorem, it is easy to show that $L^p(\Omega)$, $p > 1$ is a reflexive Banach space. Recall Definition 21.2.14 on Page 570 for the definition. From this, one obtains a weak compactness result.

23.4 Weak Compactness

Theorem 23.4.1 *For $(\Omega, \mathscr{S}, \mu)$ a σ finite measure space and $p > 1$, $L^p(\Omega)$ is reflexive. Thus every bounded sequence has a weakly convergent subsequence.*

Proof: Let $\delta_r : (L^r(\Omega))' \to L^{r'}(\Omega)$ be defined for $\frac{1}{r} + \frac{1}{r'} = 1$ by $\int (\delta_r \Lambda) g \, d\mu = \Lambda g$ for all $g \in L^r(\Omega)$. From Theorem 23.3.6 δ_r is one to one, onto, continuous and linear. By the open map theorem, δ_r^{-1} is also one to one, onto, and continuous ($\delta_r \Lambda$ equals the representor of Λ). Thus δ_r^* is also one to one, onto, and continuous by Corollary 21.2.11. Now observe that $J = \delta_p^* \circ \delta_q^{-1}$. To see this, let $z^* \in (L^q)'$, $y^* \in (L^p)'$,

$$\delta_p^* \circ \delta_q^{-1}(\delta_q z^*)(y^*) = (\delta_p^* z^*)(y^*) = z^*(\delta_p y^*) = \int (\delta_q z^*)(\delta_p y^*) d\mu,$$

$$J(\delta_q z^*)(y^*) = y^*(\delta_q z^*) = \int (\delta_p y^*)(\delta_q z^*) d\mu.$$

Therefore $\delta_p^* \circ \delta_q^{-1} = J$ on $\delta_q (L^q)' = L^p$. But the two δ maps are onto and so J is also onto. ■

What about weak compactness in $L^1(\Omega)$? I will give a simple sufficient condition in the case of a finite measure space. More can be said. See for example Dunford and Schwartz [16]. I have this in my Topics in Analysis book also. Recall Proposition 10.9.6 on Page 299 which says equi-integrable is the same as bounded and uniformly integrable. Thus in the following, you can replace equi-integrable with bounded and uniformly integrable.

Theorem 23.4.2 *Let $(\Omega, \mathscr{F}, \mu)$ be a finite measure space and let $\{f_n\}$ be a sequence in $L^1(\Omega)$ which is equi-integrable. Then there exists a subsequence which converges weakly in $L^1(\Omega)$ to some function f.*

Proof: Let

$$\mathscr{E}_n \equiv \left\{ f_n^{-1}(B(z,r)) : r \text{ is a positive rational and } z \in \mathbb{Q} + i\mathbb{Q} \right\}.$$

Let $\mathscr{E} = \cup_{n=1}^{\infty} \mathscr{E}_n$. Thus $\mathscr{E}$ and $\mathscr{E}_n$ are countable. Also, every open set is the countable union of these sets $B(z,r)$. Now let $\mathscr{K}$ be all finite intersections of sets of $\mathscr{E}$ and include $\emptyset$ and Ω in $\mathscr{K}$. Then $\sigma(\mathscr{K})$ contains inverse images of Borel sets for each f_n. Thus each f_n is measurable with respect to $\sigma(\mathscr{K})$. Also $\mathscr{K}$ is countable. Then, using a Cantor diagonal argument, we can have $\int_E g_n d\mu$ converges for all $E \in \mathscr{K}$. Let $\mathscr{G}$ be those sets $G \in \sigma(\mathscr{K})$ such that $\int_G g_n d\mu$ converges. Suppose G_k are disjoint and each in $\mathscr{G}$. Then, since Ω has finite measure, $\lim_{n\to\infty} \mu\left(\cup_{j=k}^{\infty} G_j\right) = 0$ because $\sum_k \mu(G_k)$ converges. Let $G = \cup_{k=1}^{\infty} G_k$ and so $\int_G g_n d\mu - \int_G g_m d\mu = \int_G \sum_{k=1}^{\infty} \mathscr{X}_{G_k}(g_n - g_m) d\mu = \sum_{k=1}^{N} \int_{G_k} (g_n - g_m) d\mu +$

$e(N,n,m)$ where $|e(N,n,m)| < \varepsilon$ for all n,m provided N is chosen large enough. This is by uniform integrability which is a consequence of equi-integrability. See Proposition 10.9.6. It follows that

$$\begin{aligned}\left|\int_G g_n d\mu - \int_G g_m d\mu\right| &\leq \left|\sum_{k=1}^N \int_{G_k} (g_n - g_m)\, d\mu\right| + |e(N,n,m)| \\ &< \left|\sum_{k=1}^N \int_{G_k} (g_n - g_m)\, d\mu\right| + \varepsilon < 2\varepsilon\end{aligned}$$

provided n,m are large enough. Thus $\mathscr{G}$ is closed with respect to countable disjoint unions. If $\int_G g_n d\mu$ converges, then $\int_{G^C} g_n d\mu = \int_\Omega g_n d\mu - \int_G g_n d\mu$ and so $\int_{G^C} g_n d\mu$ converges. Hence, by Dynkin's lemma, $\mathscr{G} \supseteq \sigma(\mathscr{K})$. For $E \in \sigma(\mathscr{K})$ define

$$\begin{aligned}\lambda(E) &\equiv \lim_{n\to\infty} \int_E g_n d\mu, \text{ then } \lambda \ll \mu \text{ so there is } g \text{ such that} \\ \int_E g d\mu &= \lambda(E) = \lim_{n\to\infty} \int_E g_n d\mu \text{ by Radon Nikodym, } g \in L^1\end{aligned}$$

That λ is a measure follows from the above argument that $\mathscr{G}$ is closed with respect to countable disjoint unions.

Now it was just shown that for s a simple function measurable with respect to $\sigma(\mathscr{K})$,

$$\int s g d\mu = \lim_{n\to\infty} \int s g_n d\mu.$$

Can we replace s with $h \in L^\infty(\Omega, \sigma(\mathscr{K}), \mu)$? Letting h be a representative which is uniformly bounded, there exists a sequence of simple functions $\{s_n\}$ which converges uniformly to h.

$$\left|\int h g d\mu - \int h g_n d\mu\right| \leq \left|\int h g d\mu - \int s g d\mu\right|$$

$$+ \left|\int s g d\mu - \int s g_n\right| + \left|\int s g_n d\mu - \int h g_n d\mu\right|$$

The first term on the right is no more than $\varepsilon \|g\|_{L^1}$ because s was chosen to be uniformly within ε of h. As to the last term, it is no more than $\varepsilon \max_n \|g_n\|_{L^1}$ which is no more than εC since the equi-integrability implies $\|g_n\|_{L^1}$ is bounded. The middle term converges to 0 and so $\lim_{n\to\infty} |\int h g d\mu - \int h g_n d\mu| = 0$.

Now consider $\begin{array}{ccc} L^\infty(\Omega, \sigma(\mathscr{K}), \mu) & \overset{i^*}{\leftarrow} & L^\infty(\Omega, \mathscr{F}, \mu) \\ L^1(\Omega, \sigma(\mathscr{K}), \mu) & \overset{i}{\rightarrow} & L^1(\Omega, \mathscr{F}, \mu) \end{array}$ where the inclusion map i is continuous. Let $h \in L^\infty(\Omega, \mathscr{F}, \mu)$ so $i^* h \in L^\infty(\Omega, \sigma(\mathscr{K}), \mu)$. Then

$$\lim_{n\to\infty} \int h g_n d\mu = \lim_{n\to\infty} \int h i g_n d\mu = \lim_{n\to\infty} \int i^* h g_n d\mu = \lim_{n\to\infty} \int i^* h g d\mu = \lim_{n\to\infty} \int h g d\mu$$

and this shows that g_n converges weakly to g. ■

One can extend this to an arbitrary measure space by fussing with more details that involve consideration of a σ algebra which is σ finite.

23.5 The Dual Space of $L^\infty(\Omega)$

What about the dual space of $L^\infty(\Omega)$? This will involve the following Lemma. Also recall the notion of total variation defined in Definition 23.2.2.

Lemma 23.5.1 *Let $(\Omega,\mathscr{F})$ be a measure space. Denote by $BV(\Omega)$ the space of **finitely** additive complex measures ν such that $|\nu|(\Omega)<\infty$. This means that if $\{E_i\}_{i=1}^n$ is disjoint, then $\nu\left(\cup_{i=1}^n E_i\right)=\sum_{i=1}^n \nu(E_i)$ for any $n\in\mathbb{N}$. Then defining $\|\nu\|\equiv|\nu|(\Omega)$, it follows that $BV(\Omega)$ is a Banach space.*

Proof: It is obvious that $BV(\Omega)$ is a vector space with the obvious conventions involving scalar multiplication. Why is $\|\cdot\|$ a norm? All the axioms are obvious except for the triangle inequality. However, this is not too hard either.

$$\|\mu+\nu\|\equiv|\mu+\nu|(\Omega)=\sup_{\pi(\Omega)}\left\{\sum_{A\in\pi(\Omega)}|\mu(A)+\nu(A)|\right\}$$

$$\leq\sup_{\pi(\Omega)}\left\{\sum_{A\in\pi(\Omega)}|\mu(A)|\right\}+\sup_{\pi(\Omega)}\left\{\sum_{A\in\pi(\Omega)}|\nu(A)|\right\}\equiv|\mu|(\Omega)+|\nu|(\Omega)=\|\nu\|+\|\mu\|.$$

Suppose now that $\{\nu_n\}$ is a Cauchy sequence. For each $E\in\mathscr{F}$, $|\nu_n(E)-\nu_m(E)|\leq\|\nu_n-\nu_m\|$ and so the sequence of complex numbers $\nu_n(E)$ converges. That to which it converges is called $\nu(E)$. Then it is obvious that $\nu(E)$ is finitely additive. Why is $|\nu|$ finite? Since $\|\cdot\|$ is a norm, it follows that there exists a constant C such that for all n, $|\nu_n|(\Omega)<C$. Let $\pi(\Omega)$ be any partition. Then $\sum_{A\in\pi(\Omega)}|\nu(A)|=\lim_{n\to\infty}\sum_{A\in\pi(\Omega)}|\nu_n(A)|\leq C$. Hence $\nu\in BV(\Omega)$. Let $\varepsilon>0$ be given and let N be such that if $n,m>N$, then $\|\nu_n-\nu_m\|<\varepsilon/2$. Pick any such n. Then choose $\pi(\Omega)$ such that

$$|\nu-\nu_n|(\Omega)-\varepsilon/2<\sum_{A\in\pi(\Omega)}|\nu(A)-\nu_n(A)|$$

$$=\lim_{m\to\infty}\sum_{A\in\pi(\Omega)}|\nu_m(A)-\nu_n(A)|<\lim\inf_{m\to\infty}|\nu_n-\nu_m|(\Omega)\leq\varepsilon/2$$

It follows that $\lim_{n\to\infty}\|\nu-\nu_n\|=0$. ■

Corollary 23.5.2 *Suppose $(\Omega,\mathscr{F})$ is a measure space as above and suppose μ is a measure defined on $\mathscr{F}$. Denote by $BV(\Omega;\mu)$ those finitely additive measures of $BV(\Omega)$ ν such that $\nu\ll\mu$ in the usual sense that if $\mu(E)=0$, then $\nu(E)=0$. Then $BV(\Omega;\mu)$ is a closed subspace of $BV(\Omega)$.*

Proof: It is clear that it is a subspace. Is it closed? Suppose $\nu_n\to\nu$ and each ν_n is in $BV(\Omega;\mu)$. Then if $\mu(E)=0$, it follows that $\nu_n(E)=0$ and so $\nu(E)=0$ also, being the limit of 0. ■

Definition 23.5.3 *For s a simple function $s(\omega)=\sum_{k=1}^n c_k\mathscr{X}_{E_k}(\omega)$ and $\nu\in BV(\Omega)$, define an "integral" with respect to ν as follows. $\int s d\nu\equiv\sum_{k=1}^n c_k\nu(E_k)$. For f function which is in $L^\infty(\Omega;\mu)$, define $\int f d\nu$ as follows. Applying Theorem 9.1.6, to the positive and negative parts of real and imaginary parts of f, there exists a sequence of simple functions $\{s_n\}$ which converges uniformly to f off a set of μ measure zero. Then $\int f d\nu\equiv\lim_{n\to\infty}\int s_n d\nu$*

Lemma 23.5.4 *The above definition of the integral with respect to a finitely additive measure in BV* $(\Omega;\mu)$ *is well defined.*

Proof: First consider the claim about the integral being well defined on the simple functions. This is clearly true if it is required that the c_k are disjoint and the E_k also disjoint having union equal to Ω. Thus define the integral of a simple function in this manner. First write the simple function as $\sum_{k=1}^{n} c_k \mathcal{X}_{E_k}$ where the c_k are the values of the simple function. Then use the above formula to define the integral. Next suppose the E_k are disjoint but the c_k are not necessarily distinct. Let the distinct values of the c_k be $a_1, \cdots, a_m$

$$\begin{aligned}\sum_k c_k \mathcal{X}_{E_k} &= \sum_j a_j \left(\sum_{i:c_i=a_j} \mathcal{X}_{E_i} \right) = \sum_j a_j \nu\left(\cup_{i:c_i=a_j} E_i\right) \\ &= \sum_j a_j \sum_{i:c_i=a_j} \nu(E_i) = \sum_k c_k \nu(E_k)\end{aligned}$$

and so the same formula for the integral of a simple function is obtained in this case also. Now consider two simple functions $s = \sum_{k=1}^{n} a_k \mathcal{X}_{E_k}$, $t = \sum_{j=1}^{m} b_j \mathcal{X}_{F_j}$ where the a_k and b_j are the distinct values of the simple functions. Then from what was just shown,

$$\int (\alpha s + \beta t)\, d\nu = \int \left(\sum_{k=1}^{n} \sum_{j=1}^{m} \alpha a_k \mathcal{X}_{E_k \cap F_j} + \sum_{j=1}^{m} \sum_{k=1}^{n} \beta b_j \mathcal{X}_{E_k \cap F_j} \right) d\nu$$

$$= \int \left(\sum_{j,k} \alpha a_k \mathcal{X}_{E_k \cap F_j} + \beta b_j \mathcal{X}_{E_k \cap F_j} \right) d\nu = \sum_{j,k} (\alpha a_k + \beta b_j)\, \nu(E_k \cap F_j)$$

$$\begin{aligned} &= \sum_{k=1}^{n} \sum_{j=1}^{m} \alpha a_k \nu(E_k \cap F_j) + \sum_{j=1}^{m} \sum_{k=1}^{n} \beta b_j \nu(E_k \cap F_j) \\ &= \sum_{k=1}^{n} \alpha a_k \nu(E_k) + \sum_{j=1}^{m} \beta b_j \nu(F_j) = \alpha \int s\, d\nu + \beta \int t\, d\nu \end{aligned}$$

Thus the integral is linear on simple functions so, in particular, the formula given in the above definition is well defined regardless.

So what about the definition for $f \in L^{\infty}(\Omega;\mu)$? Since $f \in L^{\infty}$, there is a set of μ measure zero N such that on N^C there exists a sequence of simple functions which converges uniformly to f on N^C. Consider s_n and s_m. As in the above, they can be written as $\sum_{k=1}^{p} c_k^n \mathcal{X}_{E_k}$, $\sum_{k=1}^{p} c_k^m \mathcal{X}_{E_k}$ respectively, where the E_k are disjoint having union equal to Ω. Then by uniform convergence, if m,n are sufficiently large, $\left|c_k^n - c_k^m\right| < \varepsilon$ or else the corresponding E_k is contained in N^C a set of ν measure 0 thanks to $\nu \ll \mu$. Hence

$$\begin{aligned}\left| \int s_n d\nu - \int s_m d\nu \right| &= \left| \sum_{k=1}^{p} (c_k^n - c_k^m)\, \nu(E_k) \right| \\ &\leq \sum_{k=1}^{p} |c_k^n - c_k^m|\, |\nu(E_k)| \leq \varepsilon \|\nu\|\end{aligned}$$

and so the integrals of these simple functions converge. Similar reasoning shows that the definition is not dependent on the choice of approximating sequence. ■

Note also that for s simple, $|\int s d\nu| \leq \|s\|_{L^\infty} |\nu|(\Omega) = \|s\|_{L^\infty} \|\nu\|$

Next the dual space of $L^\infty(\Omega;\mu)$ will be identified with $BV(\Omega;\mu)$. First here is a simple observation. Let $\nu \in BV(\Omega;\mu)$. Then define the following for $f \in L^\infty(\Omega;\mu)$.

$$T_\nu(f) \equiv \int f d\nu$$

Lemma 23.5.5 *For T_ν just defined, $|T_\nu f| \leq \|f\|_{L^\infty} \|\nu\|$.*

Proof: As noted above, the conclusion true if f is simple. Now if f is in L^∞, then it is the uniform limit of simple functions off a set of μ measure zero. Therefore, by the definition of the T_ν,

$$|T_\nu f| = \lim_{n\to\infty} |T_\nu s_n| \leq \lim\inf_{n\to\infty} \|s_n\|_{L^\infty} \|\nu\| = \|f\|_{L^\infty} \|\nu\|. \blacksquare$$

Thus each T_ν is in $(L^\infty(\Omega;\mu))'$.■

Here is the representation theorem, due to Kantorovitch, which describes the dual of $L^\infty(\Omega;\mu)$.

Theorem 23.5.6 *Let $\theta : BV(\Omega;\mu) \to (L^\infty(\Omega;\mu))'$ be given by $\theta(\nu) \equiv T_\nu$. Then θ is one to one, onto and preserves norms.*

Proof: It was shown in the above lemma that θ maps into $(L^\infty(\Omega;\mu))'$. It is obvious that θ is linear. Why does it preserve norms? From the above lemma,

$$\|\theta\nu\| \equiv \sup_{\|f\|_\infty \leq 1} |T_\nu f| \leq \|\nu\|$$

It remains to turn the inequality around. Let $\pi(\Omega)$ be a partition. Then

$$\sum_{A\in\pi(\Omega)} |\nu(A)| = \sum_{A\in\pi(\Omega)} \operatorname{sgn}(\nu(A))\,\nu(A) \equiv \int f d\nu$$

where $\operatorname{sgn}(\nu(A))$ is defined to be a complex number of modulus 1, $\operatorname{sgn}(\nu(A))\,\nu(A) = |\nu(A)|$ and

$$f(\omega) = \sum_{A\in\pi(\Omega)} \operatorname{sgn}(\nu(A))\,\mathcal{X}_A(\omega).$$

Therefore, choosing $\pi(\Omega)$ suitably, since $\|f\|_\infty \leq 1$,

$$\begin{aligned} \|\nu\| - \varepsilon &= |\nu|(\Omega) - \varepsilon \leq \sum_{A\in\pi(\Omega)} |\nu(A)| = T_\nu(f) \\ &= |T_\nu(f)| = |\theta(\nu)(f)| \leq \|\theta(\nu)\| \leq \|\nu\| \end{aligned}$$

Thus θ preserves norms. Hence it is one to one also. Why is θ onto?

Let $\Lambda \in (L^\infty(\Omega;\mu))'$. Then define

$$\nu(E) \equiv \Lambda(\mathcal{X}_E) \tag{23.7}$$

This is obviously finitely additive because Λ is linear. Also, if $\mu(E) = 0$, then $\mathcal{X}_E = 0$ in L^∞ and so $\Lambda(\mathcal{X}_E) = 0$. If $\pi(\Omega)$ is any partition of Ω, then

$$\begin{aligned} \sum_{A\in\pi(\Omega)} |\nu(A)| &= \sum_{A\in\pi(\Omega)} |\Lambda(\mathcal{X}_A)| = \sum_{A\in\pi(\Omega)} \operatorname{sgn}(\Lambda(\mathcal{X}_A))\Lambda(\mathcal{X}_A) \\ &= \Lambda\left(\sum_{A\in\pi(\Omega)} \operatorname{sgn}(\Lambda(\mathcal{X}_A))\,\mathcal{X}_A\right) \leq \|\Lambda\| \end{aligned}$$

and so $\|\nu\| \leq \|\Lambda\|$ showing that $\nu \in BV(\Omega;\mu)$. Also from 23.7, if $s = \sum_{k=1}^{n} c_k \mathcal{X}_{E_k}$ is a simple function,

$$\int s d\nu = \sum_{k=1}^{n} c_k \nu(E_k) = \sum_{k=1}^{n} c_k \Lambda\left(\mathcal{X}_{E_k}\right) = \Lambda\left(\sum_{k=1}^{n} c_k \mathcal{X}_{E_k}\right) = \Lambda(s)$$

Then letting $f \in L^{\infty}(\Omega;\mu)$, there exists a sequence of simple functions converging to f uniformly off a set of μ measure zero and so passing to a limit in the above with s replaced with s_n it follows that $\Lambda(f) = \int f d\nu$ and so θ is onto. ■

23.6 Non σ Finite Case

It turns out that for $p > 1$, you don't have to assume the measure space is σ finite. The Riesz representation theorem holds always. The proof involves the notion of uniform convexity. First recall Clarkson's inequalities. These fundamental inequalities were used to verify that $L^p(\Omega)$ is uniformly convex. More precisely, the unit ball in $L^p(\Omega)$ is uniformly convex.

Lemma 23.6.1 *Let $2 \leq p$. Then*

$$\left\| \frac{f+g}{2} \right\|_{L^p}^{p} + \left\| \frac{f-g}{2} \right\|_{L^p}^{p} \leq \frac{1}{2}\left(\|f\|_{L^p}^{p} + \|g\|_{L^p}^{p}\right)$$

Let $1 < p < 2$. then for $1/p + 1/q = 1$,

$$\left\| \frac{f+g}{2} \right\|_{L^p}^{q} + \left\| \frac{f-g}{2} \right\|_{L^p}^{q} \leq \left(\frac{1}{2}\|f\|_{L^p}^{p} + \frac{1}{2}\|g\|_{L^p}^{p} \right)^{q/p}$$

Recall the following definition of uniform convexity.

Definition 23.6.2 *A Banach space, X, is said to be uniformly convex if whenever $\|x_n\| \leq 1$ and $\left\|\frac{x_n + x_m}{2}\right\| \to 1$ as $n, m \to \infty$, then $\{x_n\}$ is a Cauchy sequence and $x_n \to x$ where $\|x\| = 1$. More precisely, for every $\varepsilon > 0$, there is a $\delta > 0$ such that if $\|x+y\| > 2-\delta$ for $\|x\|, \|y\|$ both 1, then $\|x-y\| < \varepsilon$.*

Observe that Clarkson's inequalities imply L^p is uniformly convex for all $p > 1$. Consider the harder case where $1 < p$. The other case is similar. Say $\|f\| = \|g\| = 1$ and $\|f+g\|_{L^p} > 2-\delta$. Then from the second inequality $\left(\frac{2-\delta}{2}\right)^{q} + \left\|\frac{f-g}{2}\right\|_{L^p}^{q} \leq 1$ and so

$$\|f-g\|_{L^p}^{q} \leq 2^{q}\left(1 - \left(\frac{2-\delta}{2}\right)^{q}\right) < \varepsilon$$

provided δ is small enough.

Uniformly convex spaces have a very nice property which is described in the following lemma. Roughly, this property is that any element of the dual space achieves its norm at some point of the closed unit ball.

Lemma 23.6.3 *Let X be uniformly convex and let $\phi \in X'$. Then there exists $x \in X$ such that $\|x\| = 1$, $\phi(x) = \|\phi\|$.*

Proof: There is nothing to show if $\phi = 0$ so suppose it is not. Let $\|x_n\| = 1$ and let $\phi(x_n) \to \|\phi\|$. Then as $n,m \to \infty$, $\phi\left(\frac{x_n+x_m}{2}\right) \to \|\phi\|$. Without loss of generality, we can also assume $\phi(x_n)$ is positive. Hence if m,n are large enough, then

$$\|\phi\|(1-\varepsilon) < \phi\left(\frac{x_n+x_m}{2}\right) \leq \|\phi\|\left\|\frac{x_n+x_m}{2}\right\|$$

Thus, if m,n are large enough, $\|x_n+x_m\| \geq 2(1-\varepsilon)$. It follows that $\lim_{m,n\to\infty}\|x_n+x_m\| = 2$ and so by uniform convexity, $\lim_{m,n\to\infty}\|x_n-x_m\| = 0$.Thus the sequence is a Cauchy sequence and so there is $x, \|x\| = 1$ and $x_n \to x$ so $\|\phi\| = \lim_{n\to\infty}\phi(x_n) = \phi(x)$. ■

The proof of the Riesz representation theorem will be based on the following lemma which says that if you can show a directional derivative exists, then it can be used to represent a functional in terms of this directional derivative. It is very interesting for its own sake.

Lemma 23.6.4 *(McShane) Let X be a complex normed linear space and let $\phi \in X'$. Suppose there exists $x \in X$, $\|x\| = 1$ with $\phi(x) = \|\phi\| \neq 0$. Let $y \in X$ and let $\psi_y(t) = \|x+ty\|$ for $t \in \mathbb{R}$. Suppose $\psi_y'(0)$ exists for each $y \in X$. Then for all $y \in X$,*

$$\psi_y'(0) + i\psi_{-iy}'(0) = \|\phi\|^{-1}\phi(y).$$

Proof: Suppose first that $\|\phi\| = 1$. The idea is to show that in the limit as $t \to 0$,

$$\frac{|1+t\phi(y)|-1}{t}, \frac{\|x+ty\|-\|x\|}{t}$$

act the same. The first part of the argument is devoted to showing this.

By assumption, there is x such that $\|x\| = 1$ and $\phi(x) = 1 = \|\phi\|$. Then $\phi(y-\phi(y)x) = 0$ and so

$$\phi(x+t(y-\phi(y)x)) = \phi(x)+t\phi(y)-t\phi(y)\overset{=1}{\phi(x)} = \phi(x) = 1 = \|\phi\|.$$

Therefore, $\|x+t(y-\phi(y)x)\| \geq 1$ since, from the above,

$$\|\phi\|\,\|x+t(y-\phi(y)x)\| = \|x+t(y-\phi(y)x)\| \geq \|\phi\| = 1$$

Also for small $t, |\phi(y)t| < 1$, and so $1 \leq \|x+t(y-\phi(y)x)\| = \|(1-\phi(y)t)x+ty\|$

$$\leq |1-\phi(y)t|\left\|x+\frac{t}{1-\phi(y)t}y\right\|.$$

Divide both sides by $|1-\phi(y)t|$. Using the standard formula for the sum of a geometric series,

$$1+t\phi(y)+o(t) = \frac{1}{1-t\phi(y)}$$

Therefore,

$$\frac{1}{|1-\phi(y)t|} = |1+\phi(y)t+o(t)| \leq \left\|x+\frac{t}{1-\phi(y)t}y\right\| = \|x+ty+o(t)\| \tag{23.8}$$

where $\lim_{t\to 0} o(t)(t^{-1}) = 0$. Thus, $|1+\phi(y)t| \leq \|x+ty\|+o(t)$. Now since $t\phi(y) \in \mathbb{C}$,

$$|1+t\phi(y)|-1 \geq 1+t\operatorname{Re}\phi(y)-1 = t\operatorname{Re}\phi(y).$$

Thus for $t>0$,

$$\mathrm{Re}\,\phi(y) \leq \frac{|1+t\phi(y)|-1}{t} \overset{\|x\|=1}{\leq} \frac{\|x+ty\|-\|x\|}{t}+\frac{o(t)}{t}$$

and for $t<0$,

$$\mathrm{Re}\,\phi(y) \geq \frac{|1+t\phi(y)|-1}{t} \geq \frac{\|x+ty\|-\|x\|}{t}+\frac{o(t)}{t}$$

By assumption that the directional derivative exists, and letting $t\to 0+$ and $t\to 0-$,

$$\mathrm{Re}\,\phi(y)=\lim_{t\to 0}\frac{\|x+ty\|-\|x\|}{t}=\psi'_y(0).$$

Now $\phi(y)=\mathrm{Re}\,\phi(y)+i\,\mathrm{Im}\,\phi(y)$ so $\phi(-iy)=-i(\phi(y))=-i\,\mathrm{Re}\,\phi(y)+\mathrm{Im}\,\phi(y)$ and

$$\phi(-iy)=\mathrm{Re}\,\phi(-iy)+i\,\mathrm{Im}\,\phi(-iy).$$

Hence $\mathrm{Re}\,\phi(-iy)=\mathrm{Im}\,\phi(y)$. Consequently,

$$\begin{aligned}\phi(y)&=\mathrm{Re}\,\phi(y)+i\,\mathrm{Im}\,\phi(y)=\mathrm{Re}\,\phi(y)+i\,\mathrm{Re}\,\phi(-iy)\\&=\psi'_y(0)+i\psi'_{-iy}(0).\end{aligned}$$

This proves the lemma when $\|\phi\|=1$. For arbitrary $\phi\neq 0$, let $\phi(x)=\|\phi\|,\|x\|=1$. Then from above, if $\phi_1(y)\equiv\|\phi\|^{-1}\phi(y)$, $\|\phi_1\|=1$ and so from what was just shown,

$$\phi_1(y)=\frac{\phi(y)}{\|\phi\|}=\psi'_y(0)+i\psi_{-iy}(0)\ \blacksquare$$

This shows you can represent ϕ in terms of this directional derivative.

Now here are some short observations. For $t\in\mathbb{R}$, $p>1$, and $x,y\in\mathbb{C}$, $x\neq 0$

$$\begin{aligned}\lim_{t\to 0}\frac{|x+ty|^p-|x|^p}{t}&=p\,|x|^{p-2}(\mathrm{Re}\,x\,\mathrm{Re}\,y+\mathrm{Im}\,x\,\mathrm{Im}\,y)\\&=p\,|x|^{p-2}\mathrm{Re}\,(\bar{x}y)\end{aligned}\tag{23.9}$$

Also from convexity of $f(r)=r^p$, for $|t|<1$,

$$|x+ty|^p-|x|^p\leq\|x|+|t|\,|y\|^p-|x|^p$$

$$\begin{aligned}&=\left[(1+|t|)\left(\frac{|x|+|t|\,|y|}{1+|t|}\right)\right]^p-|x|^p\leq(1+|t|)^p\,\frac{|x|^p}{1+|t|}+\frac{|t|\,|y|^p}{1+|t|}-|x|^p\\&\leq(1+|t|)^{p-1}(|x|^p+|t|\,|y|^p)-|x|^p\leq\left((1+|t|)^{p-1}-1\right)|x|^p+2^{p-1}\,|t|\,|y|^p\end{aligned}$$

Now for $f(t)\equiv(1+t)^{p-1}$, $f'(t)$ is uniformly bounded, depending on p, for $t\in[0,1]$. Hence the above is dominated by an expression of the form

$$C_p(|x|^p+|y|^p)\,|t|\tag{23.10}$$

The above lemma and uniform convexity of L^p can be used to prove a general version of the Riesz representation theorem next. Let $p>1$ and let $\eta:L^q\to(L^p)'$ be defined by

$$\eta(g)(f)=\int_\Omega gf\,d\mu.\tag{23.11}$$

Theorem 23.6.5 *(Riesz representation theorem $p>1$) The map η is 1-1, onto, continuous, and*

$$\|\eta g\| = \|g\|,\ \|\eta\| = 1.$$

Proof: Obviously η is linear. Suppose $\eta g = 0$. Then $0 = \int gf\,d\mu$ for all $f \in L^p$. Let $f = |g|^{q-2}\overline{g}$. Then $f \in L^p$and so $0 = \int |g|^q d\mu$. Hence $g = 0$ and η is one to one. That $\eta g \in (L^p)'$ is obvious from the Holder inequality. In fact, $|\eta(g)(f)| \le \|g\|_q \|f\|_p$, and so $\|\eta(g)\| \le \|g\|_q$. To see that equality holds, let $f = |g|^{q-2}\overline{g}\,\|g\|_q^{1-q}$. Then $\|f\|_p = 1$ and

$$\eta(g)(f) = \int_\Omega |g|^q d\mu\,\|g\|_q^{1-q} = \|g\|_q.$$

Thus $\|\eta\| = 1$.

It remains to show η is onto. Let $\phi \in (L^p)'$. Is $\phi = \eta g$ for some $g \in L^q$? Without loss of generality, assume $\phi \neq 0$. By uniform convexity of L^p, Lemma 23.6.3, there exists g such that $\phi g = \|\phi\|$, $g \in L^p$, $\|g\| = 1$. For $f \in L^p$, define $\phi_f(t) \equiv \int_\Omega |g+tf|^p\,d\mu$. Thus $\psi_f(t) \equiv \|g+tf\|_p \equiv \phi_f(t)^{\frac{1}{p}}$. Does $\phi'_f(0)$ exist? Let $[g=0]$ denote the set $\{x : g(x) = 0\}$.

$$\frac{\phi_f(t)-\phi_f(0)}{t} = \int \frac{(|g+tf|^p - |g|^p)}{t} d\mu$$

From 23.10, the integrand is bounded by $C_p(|f|^p + |g|^p)$. Therefore, using 23.9, the dominated convergence theorem applies and it follows $\phi'_f(0) =$

$$\lim_{t\to 0}\frac{\phi_f(t)-\phi_f(0)}{t} = \lim_{t\to 0}\left[\int_{[g=0]} |t|^{p-1}|f|^p d\mu + \int_{[g\neq 0]} \frac{(|g+tf|^p-|g|^p)}{t} d\mu\right]$$

$$= p\int_{[g\neq 0]} |g|^{p-2}\,\mathrm{Re}\,(\bar{g}f)\,d\mu = p\int |g|^{p-2}\,\mathrm{Re}\,(\bar{g}f)\,d\mu$$

Hence $\psi'_f(0) = \|g\|^{\frac{-p}{q}} \int |g(x)|^{p-2}\mathrm{Re}(g(x)\bar{f}(x))d\mu$. Note $\frac{1}{p} - 1 = -\frac{1}{q}$ Therefore,

$$\psi'_{-if}(0) = \|g\|^{\frac{-p}{q}} \int |g(x)|^{p-2}\mathrm{Re}(ig(x)\bar{f}(x))d\mu.$$

But $\mathrm{Re}(ig\bar{f}) = \mathrm{Im}(-g\bar{f})$ and so by the McShane lemma,

$$\begin{aligned}
\phi(f) &= \|\phi\|\,\|g\|^{\frac{-p}{q}} \int |g(x)|^{p-2}[\mathrm{Re}(g(x)\bar{f}(x)) + i\,\mathrm{Re}(ig(x)\bar{f}(x))]d\mu \\
&= \|\phi\|\,\|g\|^{\frac{-p}{q}} \int |g(x)|^{p-2}[\mathrm{Re}(g(x)\bar{f}(x)) + i\,\mathrm{Im}(-g(x)\bar{f}(x))]d\mu \\
&= \|\phi\|\,\|g\|^{\frac{-p}{q}} \int |g(x)|^{p-2}\overline{g}(x)f(x)d\mu.
\end{aligned}$$

This shows that $\phi = \eta(\|\phi\|\,\|g\|^{\frac{-p}{q}}|g|^{p-2}\overline{g})$ and verifies η is onto. ■

23.7 The Dual Space of $C_0(X)$

Consider the dual space of $C_0(X)$ where X is a locally compact Hausdorff space. It will turn out to be a space of measures. To show this, the following lemma will be convenient. Recall this space is defined as follows.

Definition 23.7.1 *$f \in C_0(X)$ means that for every $\varepsilon > 0$ there exists a compact set K such that $|f(x)| < \varepsilon$ whenever $x \notin K$. Recall the norm on this space is*

$$\|f\|_\infty \equiv \|f\| \equiv \sup\{|f(x)| : x \in X\}$$

Also define $C_0^+(X)$ to be the nonnegative functions of $C_0(X)$.

From the representation theorem about positive linear functionals on $C_0(X)$, we know that if Λ is such a positive linear functional, then $\Lambda f = \int_X f d\mu$. What if Λ is also continuous so that $|\Lambda f| \leq \|\Lambda\| \|f\|_\infty$?

Lemma 23.7.2 *Suppose $\Lambda : C_0(X) \to \mathbb{C}$ is a positive linear functional which is also continuous. Then if μ is the Radon measure representing Λ, it follows that $\|\Lambda\| = \mu(X)$ so in particular, μ is finite.*

Proof: From the regularity of $\mu, \mu(X) = \sup\{\mu(K) : K \subseteq X, K \text{ compact}\}$. For such a K let $K \prec f \prec X$ and so it follows that $\mu(X) = \sup\{\Lambda f : f \prec X\}$. However, $0 \leq \Lambda f \leq \|\Lambda\| \|f\|_\infty \leq \|\Lambda\|$ and so $\mu(X) \leq \|\Lambda\|$. To go the other direction, use density of $C_c(X)$ in $C_0(X)$ to obtain $f \in C_c(X)$ such that $\|f\|_\infty \leq 1$ and $\|\Lambda\| < |\Lambda f| + \varepsilon$. Since Λ is a positive linear functional, one can assume that $f \geq 0$ since otherwise the inequality could be strengthened by replacing f with its positive part f^+. Then with this f,

$$\mu(X) \leq \|\Lambda\| < |\Lambda f| + \varepsilon = \Lambda f + \varepsilon \leq \mu(X) + \varepsilon$$

and so, since ε is arbitrary, $\mu(X) = \|\Lambda\|$. ■

Next consider the case where L is in $\mathscr{L}(C_0(X), \mathbb{C})$ so it is not known to take nonnegative functions to nonnegative scalars.

Lemma 23.7.3 *Let $L \in \mathscr{L}(C_0(X), \mathbb{C})$. Then there exists $\lambda : C_0^+(X) \to [0, \infty)$ which satisfies*

$$\begin{aligned} \lambda(af + bg) &= a\lambda(f) + b\lambda(g), \text{ if } a, b \geq 0 \\ |\lambda(f)| &\leq \|L\| \|f\|_\infty \end{aligned} \tag{23.12}$$

Proof: Define, for $f \in C_0^+(X)$, $\lambda(f) \equiv \sup\{|Lg| : |g| \leq f\}$. Then the second part is obvious because

$$|\lambda(f)| = \lambda(f) \leq \sup\{\|L\| \|g\|_\infty\} \leq \|L\| \|f\|_\infty$$

Consider the first claim of 23.12. It is obvious that $\lambda(0f) = 0\lambda(f)$ from the above. If $c > 0$, why is $\lambda(cf) = c\lambda(f)$? If $|g| \leq cf$, then $\frac{1}{c}|g| \leq f$ and so $\frac{1}{c}|Lg| \leq \lambda(f)$ so $|Lg| \leq c\lambda(f)$. Taking sup for all such $g, \lambda(cf) \leq c\lambda(f)$. Thus also $\lambda(f) = \lambda\left(\frac{1}{c}cf\right) \leq \frac{1}{c}\lambda(cf)$ so $c\lambda(f) \leq \lambda(cf)$ showing that $c\lambda(f) = \lambda(cf)$ if $c \geq 0$. It remains to verify that $\lambda(f_1 + f_2) = \lambda(f_1) + \lambda(f_2)$.

For $f_j \in C_0^+(X)$, there exists $g_i \in C_0(X)$ such that $|g_i| \leq f_i$ and

$$\lambda(f_1) + \lambda(f_2) \leq |L(g_1)| + |L(g_2)| + 2\varepsilon = L(\omega_1 g_1) + L(\omega_2 g_2) + 2\varepsilon$$

$$= L(\omega_1 g_1 + \omega_2 g_2) + 2\varepsilon = |L(\omega_1 g_1 + \omega_2 g_2)| + 2\varepsilon$$

where $|g_i| \leq f_i$ and $|\omega_i| = 1$ and $\omega_i L(g_i) = |L(g_i)|$. Now

$$|\omega_1 g_1 + \omega_2 g_2| \leq |g_1| + |g_2| \leq f_1 + f_2$$

and so the above shows

$$\lambda(f_1)+\lambda(f_2)\leq\lambda(f_1+f_2)+2\varepsilon.$$

Since ε is arbitrary, $\lambda(f_1)+\lambda(f_2)\leq\lambda(f_1+f_2)$. It remains to verify the other inequality.

Now let $|g|\leq f_1+f_2$,

$$|Lg|\geq\lambda(f_1+f_2)-\varepsilon.$$

Let

$$h_i(x)=\begin{cases}\frac{f_i(x)g(x)}{f_1(x)+f_2(x)} \text{ if } f_1(x)+f_2(x)>0,\\ 0 \text{ if } f_1(x)+f_2(x)=0.\end{cases}$$

Then h_i is continuous and $h_1(x)+h_2(x)=g(x)$, $|h_i|\leq f_i$. The function h_i is clearly continuous at points x where $f_1(x)+f_2(x)>0$. The reason it is continuous at a point where $f_1(x)+f_2(x)=0$ is that at every point y where $f_1(y)+f_2(y)>0$, the top description of the function gives

$$|h_i(y)|=\left|\frac{f_i(y)\,g(y)}{f_1(y)+f_2(y)}\right|\leq|g(y)|\leq f_1(y)+f_2(y)$$

so if $|y-x|$ is small enough, $|h_i(y)|$ is also small. Then it follows from this definition of the h_i that

$$\begin{aligned}-\varepsilon+\lambda(f_1+f_2) &\leq |Lg|=|Lh_1+Lh_2|\leq|Lh_1|+|Lh_2|\\ &\leq \lambda(f_1)+\lambda(f_2).\end{aligned}$$

Since $\varepsilon>0$ is arbitrary, this shows that

$$\lambda(f_1+f_2)\leq\lambda(f_1)+\lambda(f_2)\leq\lambda(f_1+f_2)\ \blacksquare$$

λ cannot be linear because it is not defined on a vector space. However, it wants to be linear. This is the content of the above lemma. Therefore, I will call λ righteous.

23.7.1 Extending Righteous Functionals

The process of extending such a righteous functional to one which is linear is the same process used earlier with the abstract Lebesgue integral. It is just like Theorem 10.7.8 except here the functional is defined on continuous functions which are nonnegative rather than measurable nonnegative functions. The inequality of 23.12 is also preserved.

Lemma 23.7.4 *Suppose λ is a mapping which has $\lambda(f)\geq 0$ which is defined on $C_0^+(X)$ such that*

$$\lambda(af+bg)=a\lambda(f)+b\lambda(g), \tag{23.13}$$

whenever $a,b\geq 0$ and $f,g\geq 0$. Then there exists a unique extension of λ to all of $C_0(X)$, Λ such that whenever $f,g\in C_0(X)$ and $a,b\in\mathbb{C}$, it follows $\Lambda(af+bg)=a\Lambda(f)+b\Lambda(g)$. If

$$|\lambda(f)|\leq C\|f\|_\infty \tag{23.14}$$

then $|\Lambda f|\leq\lambda(|f|)\leq C\|f\|_\infty$.

Proof: There is only one possible way to extend this functional to obtain a linear functional and the arguments are identical with those of Theorem 10.7.8 so I will refer to this earlier theorem for these arguments. In particular, you must have

$$\begin{aligned}\Lambda(f) &= \Lambda(\mathrm{Re}\, f)+i\Lambda(\mathrm{Im}\, f)=\Lambda(\mathrm{Re}\, f)^{+}-\Lambda(\mathrm{Re}\, f)^{-}\\ &\quad +i\left(\Lambda(\mathrm{Im}\, f)^{+}-\Lambda(\mathrm{Im}\, f)^{-}\right)\\ &= \lambda(\mathrm{Re}\, f)^{+}-\lambda(\mathrm{Re}\, f)^{-}+i\left(\lambda(\mathrm{Im}\, f)^{+}-\lambda(\mathrm{Im}\, f)^{-}\right)\end{aligned}$$

Since the nature of the functions is different, being continuous here rather than only measurable, the only thing left is to show the claim about continuity of Λ in case of 23.14.

What of the last claim that $|\Lambda f|\le\lambda(|f|)$? Let ω have $|\omega|=1$ and $|\Lambda f|=\omega\Lambda(f)$. Since Λ is linear,

$$|\Lambda f|=\omega\Lambda(f)=\Lambda(\omega f)=\Lambda(\mathrm{Re}\,\omega f)\le\Lambda\left(\mathrm{Re}\,(\omega f)^{+}\right)$$

$$=\lambda\left(\mathrm{Re}\,(\omega f)^{+}\right)\le\lambda(|f|)\le C\|f\|_{\infty}\ \blacksquare$$

Corollary 23.7.5 *Let $L\in\mathscr{L}(C_0(X),\mathbb{C})$. Then there exists $\Lambda\in\mathscr{L}(C_0(X),\mathbb{C})$ which satisfies $\|\Lambda\|=\|L\|$ but also Λ is a positive linear functional meaning it $f\ge 0$, then $\Lambda(f)\ge 0$.*

Proof: Let λ be the righteous functional defined in Lemma 23.7.3 which satisfies $|\lambda(f)|\le\|L\|\,\|f\|_{\infty}$. Then let Λ be its extension defined in Lemma 23.7.4 which also satisfies $|\Lambda(f)|\le\|L\|\,\|f\|_{\infty}$. Then this is a positive linear functional and $\|\Lambda\|\le\|L\|$. However, from the definition of λ,

$$|Lg|\le\lambda(|g|)=\Lambda(|g|)\le\|\Lambda\|\,\|g\|_{\infty}$$

and so also $\|L\|\le\|\Lambda\|$. $\blacksquare$

23.7.2 The Riesz Representation Theorem

What follows is the Riesz representation theorem for $C_0(X)'$.

Theorem 23.7.6 *Let $L\in(C_0(X))'$ for X a locally compact Hausdorf space. Then there exists a σ algebra $\mathscr{F}$ and a finite Radon measure μ and a function $\sigma\in L^{\infty}(X,\mu)$ such that for all $f\in C_0(X)$,*

$$L(f)=\int_X f\sigma d\mu.$$

Furthermore, $\mu(X)=\|L\|$, $|\sigma|=1$ a.e. and if $\nu(E)\equiv\int_E\sigma d\mu$ then $\mu=|\nu|$.

Proof: From Corollary 23.7.5, there exists a positive linear functional Λ defined on $C_0(X)$ with $\|\Lambda\|=\|L\|$. Then let μ be the Radon measure representing Λ for which, by Lemma 23.7.2, $\mu(X)=\|\Lambda\|=\|L\|$.

For $f\in C_c(X)$, $|Lf|\le\lambda(|f|)=\Lambda(|f|)=\int_X|f|d\mu=\|f\|_{L^1(\mu)}$.Since μ is both inner and outer regular thanks to it being finite, $C_c(X)$ is dense in $L^1(X,\mu)$. (See Theorem 12.2.4 for more than is needed.) Thus there is a unique extension of L to $\tilde{L}\in\left(L^1(X,\mu)\right)'$ and

by the Riesz representation theorem for the dual of $L^1(\mu)$, there exists $\sigma \in L^\infty(\mu)$ with $\tilde{L}(f) = \int_X f\sigma d\mu$. In particular,

$$L(f) = \int_X f\sigma d\mu \text{ for } f \in C_0(X).$$

It remains to verify that $|\sigma| = 1$.

If E is measurable, the regularity of μ implies there exists a sequence of nonnegative bounded functions $f_n \in C_c(X)$ such that $f_n(x) \to \mathcal{X}_E(x)$ a.e. and in $L^1(\mu)$. Then using the dominated convergence theorem,

$$\int_E d\mu = \lim_{n\to\infty} \int_X f_n d\mu = \lim_{n\to\infty} \Lambda(f_n) \geq \lim_{n\to\infty} |Lf_n| = \lim_{n\to\infty} \left| \int_X f_n \sigma d\mu \right| = \left| \int_E \sigma d\mu \right|$$

and so if $\mu(E) > 0$, $1 \geq \left| \frac{1}{\mu(E)} \int_E \sigma d\mu \right|$ which shows from Lemma 23.2.7 that $|\sigma| \leq 1$ a.e.

But also, choosing f_1 appropriately, $\|f_1\|_\infty \leq 1, |Lf_1| + \varepsilon > \|L\| = \mu(X)$. Letting $\omega(Lf_1) = |Lf_1|, |\omega| = 1$,

$$\begin{aligned} \mu(X) &= \|L\| = \sup_{\|f\|_\infty \leq 1} |Lf| \leq |Lf_1| + \varepsilon = \omega L f_1 + \varepsilon = \int_X f_1 \omega\sigma d\mu + \varepsilon \\ &= \int_X \mathrm{Re}(f_1\omega\sigma) d\mu + \varepsilon \leq \int_X |\sigma| d\mu + \varepsilon \leq \mu(X) + \varepsilon \end{aligned}$$

and since ε is arbitrary, $\mu(X) \leq \int_X |\sigma| d\mu \leq \mu(X)$ which requires $|\sigma| = 1$ a.e. since it was shown to be no larger than 1 and if it is smaller than 1 on a set of positive measure, then the above could not hold.

If $\nu(E) \equiv \int_E \sigma d\mu$,by Corollary 23.2.9, $|\nu|(E) = \int_E |\sigma| d\mu = \int_E 1 d\mu = \mu(E)$ ■

Sometimes people write $L(f) = \int_X f d\nu \equiv \int_X f\sigma d|\nu| = \int_X f\sigma d\mu$ where $\sigma d|\nu|$ is the polar decomposition of the complex measure $\nu(E) \equiv \int_E \sigma d\mu$. Then with this convention, the above representation is $L(f) = \int_X f d\nu$, $|\nu|(X) = \|L\|$. Also note that at most one ν can represent L. If there were two of them $\nu_i, i = 1,2$, then $0 = \int_X f\sigma d|\nu_1 - \nu_2|, |\sigma| = 1$, and so it will follow that $|\nu_1 - \nu_2|(X) = 0$ because you could approximate $\bar{\sigma}$ with a sequence f_n and after using the dominated convergence theorem, you would get $|\nu_1 - \nu_2|(X) = 0$. Hence $\nu_1 = \nu_2$.

Corollary 23.7.7 *Let $L \in \mathcal{L}(C_0(X), \mathbb{C})$. Then there exists a unique complex measure ν such that for all $f \in C_0(X)$, $L(f) = \int_X f d\nu$ and $|\nu|(X) = \|L\|$.*

23.8 Exercises

1. Suppose μ is a vector measure having values in $\mathbb{R}^n$ or $\mathbb{C}^n$. Can you show that $|\mu|$ must be finite? **Hint:** You might define for each $\boldsymbol{e}_i$, one of the standard basis vectors, the real or complex measure, $\mu_{\boldsymbol{e}_i}$ given by $\mu_{\boldsymbol{e}_i}(E) \equiv \boldsymbol{e}_i \cdot \mu(E)$. Why would this approach not yield anything for an infinite dimensional normed linear space in place of $\mathbb{R}^n$? Have a look at the proof of Theorem 23.1.3.

2. The Riesz representation theorem of the L^p spaces can be used to prove a very interesting inequality. Let $r, p, q \in (1,\infty)$ satisfy $\frac{1}{r} = \frac{1}{p} + \frac{1}{q} - 1$. Then $\frac{1}{q} = 1 + \frac{1}{r} - \frac{1}{p} > \frac{1}{r}$

and so $r > q$. Let $\theta \in (0,1)$ be chosen so that $\theta r = q$. Then also $\frac{1}{r} = \left(\overbrace{1 - \frac{1}{p'}}^{1/p+1/p'=1} \right) +$ $\frac{1}{q} - 1 = \frac{1}{q} - \frac{1}{p'}$ and so $\frac{\theta}{q} = \frac{1}{q} - \frac{1}{p'}$ which implies $p'(1-\theta) = q$. Now let $f \in L^p(\mathbb{R}^n)$, $g \in L^q(\mathbb{R}^n)$, $f, g \geq 0$. Justify the steps in the following argument using what was just shown that $\theta r = q$ and $p'(1-\theta) = q$. Let $h \in L^{r'}(\mathbb{R}^n)$. $\left(\frac{1}{r} + \frac{1}{r'} = 1\right)$,

$$
\begin{aligned}
& \left| \int f * g(\boldsymbol{x}) h(\boldsymbol{x}) \, dx \right| \\
= & \left| \int \int f(\boldsymbol{y}) g(\boldsymbol{x}-\boldsymbol{y}) h(\boldsymbol{x}) \, dx dy \right| \\
\leq & \int \int |f(\boldsymbol{y})| \, |g(\boldsymbol{x}-\boldsymbol{y})|^{\theta} \, |g(\boldsymbol{x}-\boldsymbol{y})|^{1-\theta} \, |h(\boldsymbol{x})| \, dy dx \\
\leq & \int \left(\int \left(|g(\boldsymbol{x}-\boldsymbol{y})|^{1-\theta} \, |h(\boldsymbol{x})| \right)^{r'} dx \right)^{1/r'} \cdot \\
& \left(\int \left(|f(\boldsymbol{y})| \, |g(\boldsymbol{x}-\boldsymbol{y})|^{\theta} \right)^{r} dx \right)^{1/r} dy \\
\leq & \left[\int \left(\int \left(|g(\boldsymbol{x}-\boldsymbol{y})|^{1-\theta} \, |h(\boldsymbol{x})| \right)^{r'} dx \right)^{p'/r'} dy \right]^{1/p'} \\
& \left[\int \left(\int \left(|f(\boldsymbol{y})| \, |g(\boldsymbol{x}-\boldsymbol{y})|^{\theta} \right)^{r} dx \right)^{p/r} dy \right]^{1/p} \\
\leq & \left[\int \left(\int \left(|g(\boldsymbol{x}-\boldsymbol{y})|^{1-\theta} \, |h(\boldsymbol{x})| \right)^{p'} dy \right)^{r'/p'} dx \right]^{1/r'} \\
& \left[\int |f(\boldsymbol{y})|^{p} \left(\int |g(\boldsymbol{x}-\boldsymbol{y})|^{\theta r} dx \right)^{p/r} dy \right]^{1/p} \\
= & \left[\int |h(\boldsymbol{x})|^{r'} \left(\int |g(\boldsymbol{x}-\boldsymbol{y})|^{(1-\theta)p'} dy \right)^{r'/p'} dx \right]^{1/r'} \|g\|_q^{q/r} \|f\|_p \\
= & \|g\|_q^{q/r} \|g\|_q^{q/p'} \|f\|_p \|h\|_{r'} = \|g\|_q \|f\|_p \|h\|_{r'} . \qquad (23.15)
\end{aligned}
$$

Young's inequality says that

$$\|f * g\|_r \leq \|g\|_q \|f\|_p . \qquad (23.16)$$

Therefore $\|f * g\|_r \leq \|g\|_q \|f\|_p$. How does this inequality follow from the above computation? Does 23.15 continue to hold if r, p, q are only assumed to be in $[1, \infty]$? Explain. Does 23.16 hold even if r, p, and q are only assumed to lie in $[1, \infty]$?

3. Suppose $(\Omega,\mu,\mathscr{S})$ is a finite measure space and that $\{f_n\}$ is a sequence of functions which converge weakly to 0 in $L^p(\Omega)$. This means that $\int_\Omega f_n g d\mu \to 0$ for every $g \in L^{p'}(\Omega)$. Suppose also that $f_n(x) \to 0$ a.e. Show that then $f_n \to 0$ in $L^{p-\varepsilon}(\Omega)$ for every $p > \varepsilon > 0$.

4. Give an example of a sequence of functions in $L^\infty(-\pi,\pi)$ which converges weak $*$ to zero but which does not converge pointwise a.e. to zero. Convergence weak $*$ to 0 means that for every $g \in L^1(-\pi,\pi)$, $\int_{-\pi}^{\pi} g(t) f_n(t)\,dt \to 0$. **Hint:** First consider $g \in C_c^\infty(-\pi,\pi)$ and maybe try something like $f_n(t) = \sin(nt)$. Do integration by parts.

5. Let $(\Omega,\mathscr{F})$ be a measurable space and let $\lambda : \mathscr{F} \to (-\infty,\infty]$ be such that if the E_i are disjoint sets in $\mathscr{F}$ then $\lambda(\cup_i E_i) = \sum_i \lambda(E_i)$ where this sum either equals a real number or $+\infty$. The Hahn decomposition says there exist measurable sets P,N such that $P \cup N = \Omega$, $P \cap N = \emptyset$, and for each $F \subseteq P$, $\lambda(F) \geq 0$ and for each $F \subseteq N$, $\lambda(F) \leq 0$. These sets P,N are called the positive set and the negative set respectively. Show the existence of the Hahn decomposition. Also explain how this decomposition is unique in the sense that if P',N' is another Hahn decomposition, then $(P \setminus P') \cup (P' \setminus P)$ has measure zero, a similar formula holding for N,N'. When you have the Hahn decomposition, as just described, you define $\lambda^+(E) \equiv \lambda(E \cap P)$, $\lambda^-(E) \equiv -\lambda(E \cap N)$. This is sometimes called the Hahn Jordan decomposition. **Hint:** You could use similar arguments leading to Theorem 10.13.5. However, this time be sure that the Hausdorff maximality argument is applied to sets which have negative measure.

6. From Problem 5 for λ having values in $(-\infty,\infty]$ you have the Hahn Jordan decomposition for the measure λ, $\lambda^+(E) \equiv \lambda(E \cap P)$, $\lambda^-(E) \equiv -\lambda(E \cap N)$. Explain why λ^- is a finite measure. **Hint:** It is a complex measure which happens to have values in $\mathbb{R}$.

7. If $\mu : \mathscr{F} \to [0,\infty)$ is a finite measure and if $\lambda : \mathscr{F} \to (-\infty,\infty)$ is another signed measure and $\lambda \ll \mu$ meaning that if $\mu(E) = 0$, then $\lambda(E) = 0$, show that $\lambda^+, \lambda^- \ll \mu$ with both being finite measures. Explain why there exists $f \in L^1(\Omega)$ with $\lambda(E) = \int \mathscr{X}_E f d\mu$.

8. Suppose λ is like the above problem but has values in $[0,\infty]$ and μ is a finite real valued measure on $\mathscr{F}$. Suppose also that $\lambda \ll \mu$. Show there exists a measurable $f \geq 0$ such that $\lambda(E) = \int \mathscr{X}_E f d\mu$

9. What if λ has values in $[-\infty,\infty)$. Prove there exists a Hahn decomposition for λ as in the above problem. Why do we not allow λ to have values in $[-\infty,\infty]$? **Hint:** You might want to consider $-\lambda$.

10. Suppose X is a Banach space and let X' denote its dual space. A sequence $\{x_n^*\}_{n=1}^\infty$ in X' is said to converge weak $*$ to $x^* \in X'$ if for every $x \in X$, $\lim_{n\to\infty} x_n^*(x) = x^*(x)$. Let $\{\phi_n\}$ be a mollifier. Also let δ be the measure defined by $\delta(E) = 1$ if $\mathbf{0} \in E$ and 0 if $1 \notin E$. Explain how $\phi_n \to \delta$ weak $*$.

11. It was shown above that if $\phi \in X'$ where X is a uniformly convex Banach space, then there exists $x \in X$, $\|x\| = 1$, and $\phi(x) = \|\phi\|$. Show that this x must be unique. **Hint:** Recall that uniform convexity implies strict convexity.

12. Suppose $\lambda(E) = \int_E h d\mu$ where h is real valued and μ is a finite measure so that λ is also real valued. Let P,N be a Hahn decomposition for λ. Show that $|\lambda|(E) = \int_E |h| d\mu$. **Hint:** Argue that on P it follows $h \geq 0$ a.e. and on $N, h \leq 0$ a.e. Then estimate $\sum_{F \in \pi(E)} \lambda(F)$ using a Hahn decomposition. If we defined $|x+iy|_1$ as $|x|_1 + |y|_1$, and the total variation exactly the same way for a complex valued measure except for letting $|\cdot|_1$ refer to this way of measuring magnitude, then everything would be much easier. Why don't we do this and save a lot of trouble?

Chapter 24

The Bochner Integral

From my experience, the Bochner integral tends to be ignored. However, it is one of the most useful and important ideas in functional analysis, at least in my experience. Perhaps it is not useful in number theory or modern algebra, but I have used it in almost every paper I have written durring my career. Therefore, I am including it in this book. If people are not interested in it, they can ignore it, but if so, they will be missing out on some very nice mathematics. The work of Pettis about weak and strong measurability is particularly interesting.

24.1 Strong and Weak Measurability

In this section $(\Omega, \mathscr{F})$ will be a measurable space and X will be a Banach space which contains the values of either a function or a measure. The Banach space will be either a real or a complex Banach space but the field of scalars does not matter and so it is denoted by $\mathbb{F}$ with the understanding that $\mathbb{F} = \mathbb{C}$ unless otherwise stated. The theory presented here includes the case where $X = \mathbb{R}^n$ or $\mathbb{C}^n$ but it does not include the situation where f could have values in something like $[0,\infty]$ which is not a vector space. To begin with here is a definition.

Definition 24.1.1 *A function, $x : \Omega \to X$, for X a Banach space, is finitely valued and measurable if it is of the form $x(\omega) = \sum_{i=1}^n a_i \mathscr{X}_{B_i}(\omega)$ where $B_i \in \mathscr{F}$ for each i. A function x from Ω to X is said to be strongly measurable if there exists a sequence of finitely valued and measurable functions $\{x_n\}$ with $x_n(\omega) \to x(\omega)$. The function x is said to be weakly measurable if, for each $f \in X'$, $f \circ x$ is a scalar valued measurable function.*

The approximating simple functions can be modified so that the norm of each is no more than $2\|x(\omega)\|$. This is a useful observation.

Lemma 24.1.2 *Let x be strongly measurable. Then $\|x\|$ is a real valued measurable function. There exists a sequence of simple functions $\{y_n\}$ which converges to $x(\omega)$ pointwise and also $\|y_n(\omega)\| \leq 2\|x(\omega)\|$ for all ω.*

Proof: Consider the first claim. Letting x_n be a sequence of simple functions converging to x pointwise, it follows that $\|x_n\|$ is a real valued measurable function. Since $\|x\|$ is a pointwise limit, $\|x\|$ is a real valued measurable function.

Let $x_n(\omega)$ be simple functions converging to $x(\omega)$ pointwise as above. Let

$$x_n(\omega) \equiv \sum_{k=1}^{m_n} a_k^n \mathscr{X}_{E_k^n}(\omega)$$

Then

$$y_n(\omega) \equiv \begin{cases} x_n(\omega) \text{ if } \|x_n(\omega)\| < 2\|x(\omega)\| \\ 0 \text{ if } \|x_n(\omega)\| \geq 2\|x(\omega)\| \end{cases}$$

Thus, for $\left[\left\|a_k^n\right\| \leq 2\|x\|\right] \equiv \left\{\omega : \left\|a_k^n\right\| \leq 2\|x(\omega)\|\right\}$,

$$y_n(\omega) = \sum_{k=1}^{m_n} a_k^n \mathscr{X}_{E_k^n \cap \left[\left\|a_k^n\right\| \leq 2\|x\|\right]}(\omega)$$

It follows y_n is a simple function. If $\|x(\omega)\| = 0$, then $y_n(\omega) = 0$ and so $y_n(\omega) \to x(\omega)$. If $\|x(\omega)\| > 0$, then eventually, $y_n(\omega) = x_n(\omega)$ and so in this case, $y_n(\omega) \to x(\omega)$. ■

Earlier, a function was measurable if inverse images of open sets were measurable. Something similar holds here. The difference is that another condition needs to hold about the values being separable. First is a somewhat obvious lemma.

Lemma 24.1.3 *Suppose S is a nonempty subset of a metric space (X,d) and $S \subseteq T$ where T is separable. Then there exists a countable dense subset of S.*

Proof: Let D be the countable dense subset of T. Now consider the countable set $\mathscr{B}$ of balls having center at a point of D and radius a positive rational number such that also, each ball in $\mathscr{B}$ has nonempty intersection with S. Let $\mathscr{D}$ consist of a point from $S \cap B$ whenever $B \in \mathscr{B}$ (axiom of choice). Let $s \in S$ and consider $B(s,\varepsilon)$. Let r be rational with $r < \varepsilon$. Now $B\left(s, \frac{r}{10}\right)$ contains a point $d \in D$. Thus $B\left(d, \frac{r}{10}\right) \in \mathscr{B}$ and $s \in B\left(d, \frac{r}{10}\right)$. Let $\hat{d} \in \mathscr{D} \cap B\left(d, \frac{r}{10}\right)$. Thus $d\left(s,\hat{d}\right) < \frac{r}{5} < r < \varepsilon$ so $\hat{d} \in B(s,\varepsilon)$ and this shows that $\mathscr{D}$ is a countable dense subset of S as claimed. ■

The following is a general result in metric space.

Lemma 24.1.4 *Let X be a metric space and suppose V is an open set in V. Then there exists open sets V_m such that*

$$\cdots V_m \subseteq \overline{V}_m \subseteq V_{m+1} \subseteq \cdots, \ V = \bigcup_{m=1}^{\infty} V_m. \tag{24.1}$$

Proof: Recall that if S is a nonempty set, $x \to \operatorname{dist}(x,S)$ is a continuous map from X to $\mathbb{R}$. First assume $V \neq X$. Let $V_m \equiv \left\{x \in V : \operatorname{dist}\left(x, V^C\right) > \frac{1}{m}\right\}$. Then for large enough m, this set is nonempty and contained in V. Furthermore, if $x \in V$ then it is at a positive distance to the closed set V^C so eventually, $x \in V_m$. Now

$$V_m \subseteq \overline{V_m} \subseteq \left\{x \in V : \operatorname{dist}\left(x, V^C\right) \geq \frac{1}{m}\right\} \subseteq V$$

Indeed, if p is a limit point of V_m, then there are $x_n \in V_m$ with $x_n \to p$. Thus $\operatorname{dist}\left(x_n, V^C\right) \to \operatorname{dist}\left(p, V^C\right)$ and so p is in the set on the right. In case $X = V$, let $V_m \equiv B(\xi, m)$. Then $V_m \subseteq \overline{V_m} \subseteq \left\{x \in V : \operatorname{dist}\left(x, V^C\right) \geq \frac{1}{m}\right\}$ and the union of these V_m equals V. ■

Theorem 24.1.5 *x is strongly measurable if and only if $x^{-1}(U)$ is measurable for all U open in X and $x(\Omega)$ is separable. Thus, if X is separable, x is strongly measurable if and only if $x^{-1}(U)$ is measurable for all U open.*

Proof: $\Leftarrow$Suppose first $x^{-1}(U)$ is measurable for all U open in X and $x(\Omega)$ is separable. It follows $x^{-1}(B)$ is measurable for all B Borel because $\{B : x^{-1}(B) \text{ is measurable}\}$ is a σ algebra containing the open sets. Let $\{a_n\}_{n=1}^{\infty}$ be the dense subset of $x(\Omega)$. Let

$$U_k^n \equiv \{z \in X : \|z - a_k\| \leq \min\{\{\|z - a_l\|\}_{l=1}^n\}.$$

In words, U_k^m is the set of points of X which are as close to a_k as they are to any of the a_l for $l \leq n$.

$$B_k^n \equiv x^{-1}(U_k^n), \ D_k^n \equiv B_k^n \setminus \left(\cup_{i=1}^{k-1} B_i^n\right), \ D_1^n \equiv B_1^n,$$

and $x_n(\omega) \equiv \sum_{k=1}^n a_k \mathcal{X}_{D_k^n}(\omega)$. Thus $x_n(\omega)$ is a closest approximation to $x(\omega)$ from $\{a_k\}_{k=1}^n$ and so $x_n(\omega) \to x(\omega)$ because $\{a_n\}_{n=1}^\infty$ is dense in $x(\Omega)$. Furthermore, x_n is measurable because each D_k^n is measurable.

$\Rightarrow$Now suppose that x is strongly measurable. Then some sequence of measurable finite valued functions $\{x_n\}$ converges pointwise to x. Then $x_n^{-1}(W)$ is measurable for every open set W because it is just a finite union of measurable sets. If $x_n(\omega) = \sum_{k=1}^n c_k \mathcal{X}_{E_k}(\omega)$,

$$x_n^{-1}(W) = \cup\{E_k : c_k \in W\}$$

Thus, $x_n^{-1}(W)$ is measurable for every Borel set W. This follows as before since

$$\left\{W : x_n^{-1}(W) \text{ is measurable}\right\}$$

is a σ algebra containing the open sets. Since X is a metric space, it follows that if U is an open set in X, there exists a sequence of open sets, $\{V_n\}$ which satisfies

$$\overline{V}_n \subseteq U,\ \overline{V}_n \subseteq V_{n+1},\ U = \cup_{n=1}^\infty V_n.$$

Then $x^{-1}(V_m) \subseteq \bigcup\limits_{n<\infty} \bigcap\limits_{k\geq n} x_k^{-1}(V_m) \subseteq x^{-1}\left(\overline{V}_m\right)$. This implies $x^{-1}(U) = \bigcup\limits_{m<\infty} x^{-1}(V_m)$

$$\subseteq \bigcup_{m<\infty} \bigcup_{n<\infty} \bigcap_{k\geq n} x_k^{-1}(V_m) \subseteq \bigcup_{m<\infty} x^{-1}\left(\overline{V}_m\right) \subseteq x^{-1}(U).$$

Since $x^{-1}(U) = \bigcup_{m<\infty} \bigcup\limits_{n<\infty} \bigcap\limits_{k\geq n} x_k^{-1}(V_m)$, it follows that $x^{-1}(U)$ is measurable for every open U. It remains to show $x(\Omega)$ is separable. Let $D \equiv$ all values of the x_n. Then $x(\Omega) \subseteq \overline{D}$, which has a countable dense subset. By Lemma 24.1.3, $x(\Omega)$ is separable. ■

Lemma 24.1.6 *Let $x \in X$ a normed linear space. Then there exists $f \in X'$ such that $\|f\| = 1$ and $f(x) = \|x\|$.*

Proof: Consider the one dimensional subspace

$$M \equiv \left\{\alpha \frac{x}{\|x\|} : \alpha \in \mathbb{F}\right\}$$

and define a continuous linear functional on M by $g\left(\alpha \frac{x}{\|x\|}\right) \equiv \alpha$. Then the norm of $\|g\| \equiv \sup_{|\alpha|\leq 1} |\alpha| = 1$. Extend g to all of X using the Hahn Banach theorem calling the extended function f. Then $\|f\| = 1$ and $f(x) = f\left(\|x\| \frac{x}{\|x\|}\right) = \|x\|$. ■

The next lemma is interesting for its own sake. Roughly it says that if a Banach space is separable, then the unit ball in the dual space is weak $*$ separable. This will be used to prove Pettis's theorem, one of the major theorems in this subject which relates weak measurability to strong measurability. First here is a standard application which comes from earlier material on the Hahn Banach theorem.

Lemma 24.1.7 *If X is a separable Banach space with B' the closed unit ball in X', then there exists a sequence $\{f_n\}_{n=1}^\infty \equiv D' \subseteq B'$ with the property that for every $x \in X$,*

$$\|x\| = \sup_{f\in D'} |f(x)|$$

If H is a dense subset of X' then D' may be chosen to be contained in H.

Proof: Let $\{a_k\}_{k=1}^{\infty}$ be a countable dense set in X and consider the mapping

$$\phi_n : B' \to \mathbb{F}^n$$

given by

$$\phi_n(f) \equiv (f(a_1), \cdots, f(a_n)).$$

Then $\phi_n(B')$ is contained in a compact subset of $\mathbb{F}^n$ because $|f(a_k)| \leq \|a_k\|$. Therefore, there exists a countable dense subset of $\phi_n(B'), \{\phi_n(f_k)\}_{k=1}^{\infty}$. Pick $h_j^k \in H \cap B'$ such that $\lim_{j\to\infty} \left\| f_k - h_j^k \right\| = 0$. Then $\left\{ \phi_n\left(h_j^k\right), k, j \right\}$ must also be dense in $\phi_n(B')$. Let $D_n' = \left\{ h_j^k, k, j \right\}$. Thus D_n' is a countable collection of $f \in B'$ which can be used to approximate each $\|a_k\|, k \leq n$. Indeed, if x is arbitrary, there exists $f_x \in B'$ with $f_x(x) = \|x\|$. Thus $\|a_k\|$ is contained in $\phi_n(B')$. Define

$$D' \equiv \cup_{n=1}^{\infty} D_n'.$$

From the construction, D' is countable and can be used to approximate each $\|a_m\|$. That is,

$$\|a_m\| = \sup\{|f(a_m)| : f \in D'\}$$

Then, for x arbitrary, $|f(x)| \leq \|x\|$ and so

$$\begin{aligned} \|x\| &\leq \|x - a_m\| + \|a_m\| = \|x - a_m\| + \sup\{|f(a_m)| : f \in D'\} \\ &\leq \sup\{|f(a_m - x) + f(x)| : f \in D'\} + \|x - a_m\| \\ &\leq \sup\{|f(x)| : f \in D'\} + 2\|x - a_m\| \leq \|x\| + 2\|x - a_m\|. \end{aligned}$$

Since a_m is arbitrary and the $\{a_m\}_{m=1}^{\infty}$ are dense, this establishes the claim of the lemma. ■

Note that the proof would work the same if H were only given to be weak $*$ dense.

The next theorem is one of the most important results in the subject. It is due to Pettis and appeared in 1938 [44].

Theorem 24.1.8

If x has values in a separable Banach space X, then x is weakly measurable if and only if x is strongly measurable.

Proof: $\Rightarrow$It is necessary to show $x^{-1}(U)$ is measurable whenever U is open. Since every open set is a countable union of balls, it suffices to show $x^{-1}(B(a,r))$ is measurable for any ball, $B(a,r)$. Since every open ball is the countable union of closed balls, it suffices to verify $x^{-1}\left(\overline{B(a,r)}\right)$ is measurable. For D' described in Lemma 24.1.7,

$$\begin{aligned} x^{-1}\left(\overline{B(a,r)}\right) &= \{\omega : \|x(\omega) - a\| \leq r\} = \left\{ \omega : \sup_{f \in D'} |f(x(\omega) - a)| \leq r \right\} \\ &= \cap_{f \in D'} \{\omega : |f(x(\omega) - a)| \leq r\} \\ &= \cap_{f \in D'} \{\omega : |f(x(\omega)) - f(a)| \leq r\} \\ &= \cap_{f \in D'} (f \circ x)^{-1} \overline{B(f(a), r)} \end{aligned}$$

which equals a countable union of measurable sets because it is assumed that $f \circ x$ is measurable for all $f \in X'$.

$\Leftarrow$Next suppose x is strongly measurable. Then there exists a sequence of simple functions x_n which converges to x pointwise. Hence for all $f \in X'$, $f \circ x_n$ is measurable and $f \circ x_n \to f \circ x$ pointwise. Thus x is weakly measurable. ■

The same method of proof yields the following interesting corollary.

Corollary 24.1.9 *Let X be a separable Banach space and let $\mathscr{B}(X)$ denote the σ algebra of Borel sets. Let H be a dense subset of X'. Then $\mathscr{B}(X) = \sigma(H) \equiv \mathscr{F}$, where $\sigma(H)$ is the smallest σ algebra of subsets of X which has the property that every function, $x^* \in H$ is measurable.*

Proof: First I need to show $\mathscr{F}$ contains open balls because then $\mathscr{F}$ will contain the open sets and hence the Borel sets. As noted above, it suffices to show $\mathscr{F}$ contains closed balls. Let D' be those functionals in B' defined in Lemma 24.1.7 contained in H. Then

$$\begin{aligned}
\{x : \|x-a\| \le r\} &= \left\{x : \sup_{x^* \in D'} |x^*(x-a)| \le r\right\} \\
&= \cap_{x^* \in D'} \{x : |x^*(x-a)| \le r\} \\
&= \cap_{x^* \in D'} \{x : |x^*(x) - x^*(a)| \le r\} \\
&= \cap_{x^* \in D'} x^{*-1}\left(\overline{B(x^*(a), r)}\right) \in \sigma(H)
\end{aligned}$$

which is measurable because this is a countable intersection of measurable sets. Thus $\mathscr{F}$ contains open sets so $\sigma(H) \equiv \mathscr{F} \supseteq \mathscr{B}(X)$.

To show the other direction for the inclusion, note that each x^* is $\mathscr{B}(X)$ measurable because $x^{*-1}(\text{open set}) = \text{open set}$. Therefore, $\mathscr{B}(X) \supseteq \sigma(H)$. ■

What of limits of measurable functions? The next theorem says that the usual theorem about limits of measurable functions being measurable holds. The proof is similar to showing that the limit of measurable finitely valued functions is measurable given above.

Theorem 24.1.10 *Let x_n and x be functions mapping Ω to X where $\mathscr{F}$ is a σ algebra of measurable sets of Ω and X is a Banach space. Thus X satisfies 24.1. Then if x_n is strongly measurable, and $x(\omega) = \lim_{n\to\infty} x_n(\omega)$, it follows that x is also strongly measurable. (Pointwise limits of measurable functions are measurable.)*

Proof: Let $\{V_m\}$ be the sequence of 24.1. Since x is the pointwise limit of x_n,

$$x^{-1}(V_m) \subseteq \{\omega : x_k(\omega) \in V_m \text{ for all } k \text{ large enough}\} \subseteq x^{-1}(\overline{V_m}).$$

Therefore,

$$\begin{aligned}
x^{-1}(V) = \cup_{m=1}^{\infty} x^{-1}(V_m) &\subseteq \cup_{m=1}^{\infty} \cup_{n=1}^{\infty} \cap_{k=n}^{\infty} x_k^{-1}(V_m) \\
&\subseteq \cup_{m=1}^{\infty} x^{-1}(\overline{V_m}) = x^{-1}(V).
\end{aligned}$$

It follows $x^{-1}(V) \in \mathscr{F}$ because it equals the expression in the middle which is measurable. Note that this shows the characterization of measurability in terms of inverse images of open sets being measureable sets. Thus the theorem is proved in the case of separable Banach spaces. However, Lemma 24.1.3 can be applied to conclude that this holds in general because each x_n is separably valued given they are each strongly measurable and $x(\Omega) \subseteq \overline{D}$ where $D = \cup_n D_n$ for D_n a countable dense subset of $x_n(\Omega)$. ■

Note that the same conclusion in terms of inverse images being measurable would hold for any metric space.

Corollary 24.1.11 *x is strongly measurable if and only if $x(\Omega)$ is separable and x is weakly measurable.*

Proof: Strong measurability clearly implies weak measurability. If $x_n(\omega) \to x(\omega)$ where x_n is simple, then $f(x_n(\omega)) \to f(x(\omega))$ for all $f \in X'$. Hence $f \circ x$ is measurable by Theorem 24.1.10 because it is the limit of a sequence of measurable functions. Let D denote the set of all values of x_n. Then $\overline{D}$ is a separable set containing $x(\Omega)$. Thus $\overline{D}$ is a separable metric space. Therefore $x(\Omega)$ is separable also by the last part of the proof of Theorem 24.1.5.

Now suppose D is a countable dense subset of $x(\Omega)$ and x is weakly measurable. Let Z be the subset consisting of all finite linear combinations of D with the scalars coming from the set of rational points of $\mathbb{F}$. Thus, Z is countable. Letting $Y = \overline{Z}$, Y is a separable Banach space containing $x(\Omega)$. If $f \in Y'$, f can be extended to an element of X' by the Hahn Banach theorem. Therefore, x is a weakly measurable Y valued function. Now use Theorem 24.1.8 to conclude x is strongly measurable. ■

Weakly measurable as defined above means $\omega \to x^*(x(\omega))$ is measurable for every $x^* \in X'$. The next lemma ties this weak measurability to the usual version of measurability in which a function is measurable when inverse images of open sets are measurable.

Lemma 24.1.12 *Let X be a Banach space and let $x : (\Omega, \mathscr{F}) \to K \subseteq X$ where K is weakly compact and X' is separable. Then x is weakly measurable if and only if $x^{-1}(U) \in \mathscr{F}$ whenever U is a weakly open set.*

Proof: By Corollary 21.5.11 on Page 583, there exists a metric d, such that the metric space topology with respect to d coincides with the weak topology on K. Since K is compact, it follows that K is also separable. Hence it is completely separable and so there exists a countable basis of open sets $\mathscr{B}$ for the weak topology on K. It follows that if U is any weakly open set, covered by basic sets of the form $B_A(x,r)$ where A is a finite subset of X', there exists a countable collection of these sets of the form $B_A(x,r)$ which covers U.

Suppose now that x is weakly measurable. To show $x^{-1}(U) \in \mathscr{F}$ whenever U is weakly open, it suffices to verify $x^{-1}(B_A(z,r)) \in \mathscr{F}$ for any set, $B_A(z,r)$. Let $A = \{x_1^*, \cdots, x_m^*\}$. Then

$$\begin{aligned} x^{-1}(B_A(z,r)) &= \{\omega \in \Omega : \rho_A(x(\omega) - z) < r\} \\ &\equiv \left\{\omega \in \Omega : \max_{x^* \in A} |x^*(x(\omega) - z)| < r\right\} \\ &= \cup_{i=1}^m \{\omega \in \Omega : |x_i^*(x(\omega) - z)| < r\} \\ &= \cup_{i=1}^m \{\omega \in \Omega : |x_i^*(x(\omega)) - x_i^*(z)| < r\} \end{aligned}$$

which is measurable because each $x_i^* \circ x$ is given to be measurable.

Next suppose $x^{-1}(U) \in \mathscr{F}$ whenever U is weakly open. Then in particular this holds when $U = B_{x^*}(z,r)$ for arbitrary x^*. Hence

$$\{\omega \in \Omega : x(\omega) \in B_{x^*}(z,r)\} \in \mathscr{F}.$$

But this says the same as

$$\{\omega \in \Omega : |x^*(x(\omega)) - x^*(z)| < r\} \in \mathscr{F}$$

Since $x^*(z)$ can be a completely arbitrary element of $\mathbb{F}$, it follows $x^* \circ x$ is an $\mathbb{F}$ valued measurable function. In other words, x is weakly measurable according to the former definition. ■

One can also define weak $*$ measurability and prove a theorem just like the Pettis theorem above. The next lemma is the analogue of Lemma 24.1.7.

Lemma 24.1.13 *Let B be the closed unit ball in X. If X' is separable, there exists a sequence $\{x_m\}_{m=1}^{\infty} \equiv D \subseteq B$ with the property that for all $y^* \in X'$,*

$$\|y^*\| = \sup_{x \in D} |y^*(x)|.$$

Proof: Let $\{x_k^*\}_{k=1}^{\infty}$ be the dense subset of X'. Define $\phi_n : B \to \mathbb{F}^n$ by

$$\phi_n(x) \equiv (x_1^*(x), \cdots, x_n^*(x)).$$

Then $|x_k^*(x)| \leq \|x_k^*\|$ and so $\phi_n(B)$ is contained in a compact subset of $\mathbb{F}^n$. Therefore, there exists a countable set, $D_n \subseteq B$ such that $\phi_n(D_n)$ is dense in $\phi_n(B)$. That is, $\{(x_1^*(x), \cdots, x_n^*(x)) : x \in D_n\}$ is dense in $\phi_n(B)$.

$$D \equiv \cup_{n=1}^{\infty} D_n.$$

It remains to verify this works. Let $y^* \in X'$. I want to show that $\|y^*\| = \sup_{x \in D} |y^*(x)|$. There exists $y, \|y\| \leq 1$, such that

$$|y^*(y)| > \|y^*\| - \varepsilon.$$

By density, there exists one of the x_k^* from the countable dense subset of X' such that also

$$\|x_k^* - y^*\| < \varepsilon, \text{ so } |x_k^*(y)| > \|y^*\| - 2\varepsilon$$

Now $x_k^*(y) \in \phi_k(B)$ and so there exists $x \in D_k \subseteq D \subseteq B$ such that also

$$|x_k^*(x)| > \|y^*\| - 2\varepsilon.$$

Then since $\|x_k^* - y^*\| < \varepsilon$, this implies

$$\|y^*\| \geq |y^*(x)| = |(y^* - x_k^*)(x) + x_k^*(x)| \geq |x_k^*(x)| - \varepsilon > \|y^*\| - 3\varepsilon$$

It follows that

$$\|y^*\| - 3\varepsilon \leq \sup_{x \in D} |y^*(x)| \leq \|y^*\|$$

This proves the lemma because ε is arbitrary. ■

The next theorem is another version of the Pettis theorem. First here is a definition.

Definition 24.1.14 *A function y having values in X' is weak $*$ measurable, when for each $x \in X$, $y(\cdot)(x)$ is a measurable scalar valued function.*

Theorem 24.1.15 *If X' is separable and $y : \Omega \to X'$ is weak $*$ measurable meaning $\omega \to y(\omega)(x)$ is a $\mathbb{F}$ valued measurable function, then y is strongly measurable.*

Proof: It is necessary to show $y^{-1}(B(a^*,r))$ is measurable for $a^* \in X'$. This will suffice because the separability of X' implies every open set is the countable union of such balls of the form $B(a^*,r)$. It also suffices to verify inverse images of closed balls are measurable because every open ball is the countable union of closed balls. From Lemma 24.1.13,

$$\begin{aligned} y^{-1}\left(\overline{B(a^*,r)}\right) &= \{\omega : \|y(\omega) - a^*\| \leq r\} \\ &= \left\{\omega : \sup_{x\in D} |(y(\omega) - a^*)(x)| \leq r\right\} \\ &= \left\{\omega : \sup_{x\in D} |y(\omega)(x) - a^*(x)| \leq r\right\} \\ &= \cap_{x\in D} y(\cdot)(x)^{-1}\left(\overline{B(a^*(x),r)}\right) \end{aligned}$$

which is a countable intersection of measurable sets by hypothesis. ■

The following are interesting consequences of the theory developed so far and are of interest independent of the theory of integration of vector valued functions.

Theorem 24.1.16 *If X' is separable, then so is X.*

Proof: Let $D = \{x_m\} \subseteq B$, the unit ball of X, be the sequence promised by Lemma 24.1.13. Let V be all finite linear combinations of elements of $\{x_m\}$ with rational scalars. Thus $\overline{V}$ is a separable subspace of X. The claim is that $\overline{V} = X$. If not, there exists

$$x_0 \in X \setminus \overline{V}.$$

But by the Hahn Banach theorem there exists $x_0^* \in X'$ satisfying $x_0^*(x_0) \neq 0$, but $x_0^*(v) = 0$ for every $v \in \overline{V}$. Hence

$$\|x_0^*\| = \sup_{x\in D} |x_0^*(x)| = 0,$$

a contradiction. ■

Corollary 24.1.17 *If X is reflexive, then X is separable if and only if X' is separable.*

Proof: From the above theorem, if X' is separable, then so is X. Now suppose X is separable with a dense subset equal to D. Then since X is reflexive, $J(D)$ is dense in X'' where J is the James map satisfying $Jx(x^*) \equiv x^*(x)$. Then since X'' is separable, it follows from the above theorem that X' is also separable. ■

Note how this shows that $L^1(\mathbb{R}^p, m_p)$ is not reflexive because this is a separable space, but $L^\infty(\mathbb{R}^p, m_p)$ is clearly not. For example, you could consider $\mathcal{X}_{[0,r]}$ for r a positive irrational number. There are uncountably many of these functions in $L^\infty([0,1])$ and it is clear that $\left\|\mathcal{X}_{[0,r]} - \mathcal{X}_{[0,\hat{r}]}\right\|_\infty = 1$.

24.1.1 Eggoroff's Theorem

In the context of a more general notion of measurable function having values in a metric space, here is a version of Egoroff's theorem. Here we introduce a finite measure μ. None of the above section had anything to do with a measure.

Theorem 24.1.18 *(Egoroff) Let $(\Omega,\mathscr{F},\mu)$ be a finite measure space,$(\mu(\Omega)<\infty)$ and let f_n, f be X valued measurable functions where X is a separable metric space and for all $\omega\notin E$ where $\mu(E)=0, f_n(\omega)\to f(\omega)$. Then for every $\varepsilon>0$, there exists a set, $F\supseteq E$, $\mu(F)<\varepsilon$, such that f_n converges uniformly to f on F^C.*

Proof: First suppose $E=\emptyset$ so that convergence is pointwise everywhere. Let

$$E_{km}=\{\omega\in\Omega : d(f_n(\omega),f(\omega))\geq 1/m \text{ for some } n>k\}.$$

Claim: $\left[\omega : d(f_n(\omega),f(\omega))\geq \frac{1}{m}\right]$ is measurable.

Proof of claim: Let $\{x_k\}_{k=1}^{\infty}$ be a countable dense subset of X and let r denote a positive rational number, $\mathbb{Q}^+$. Then

$$\cup_{k\in\mathbb{N},r\in\mathbb{Q}^+} f_n^{-1}(B(x_k,r))\cap f^{-1}\left(B\left(x_k,\frac{1}{m}-r\right)\right)=\left[d(f,f_n)<\frac{1}{m}\right] \tag{24.2}$$

Here is why. If ω is in the set on the left, then $d(f_n(\omega),x_k)<r$ and $d(f(\omega),x_k)<\frac{1}{m}-r$. Therefore,

$$d(f(\omega),f_n(\omega))<r+\frac{1}{m}-r=\frac{1}{m}.$$

Thus the left side is contained in the right. Now let ω be in the right side. That is $d(f_n(\omega),f(\omega))<\frac{1}{m}$. Choose $2r<\frac{1}{m}-d(f_n(\omega),f(\omega))$ and pick $x_k\in B(f_n(\omega),r)$. Then

$$d(f(\omega),x_k)\leq d(f(\omega),f_n(\omega))+d(f_n(\omega),x_k)<\frac{1}{m}-2r+r=\frac{1}{m}-r$$

Thus $\omega\in f_n^{-1}(B(x_k,r))\cap f^{-1}\left(B\left(x_k,\frac{1}{m}-r\right)\right)$ and so ω is in the left side. Thus the two sets are equal. Now the set on the left in 24.2 is measurable because it is a countable union of measurable sets. This proves the claim since $\left[\omega : d(f_n(\omega),f(\omega))\geq\frac{1}{m}\right]$ is the complement of this measurable set.

Hence E_{km} is measurable because $E_{km}=\cup_{n=k+1}^{\infty}\left[\omega : d(f_n(\omega),f(\omega))\geq\frac{1}{m}\right]$. For fixed $m,\cap_{k=1}^{\infty}E_{km}=\emptyset$ because $f_n(\omega)$ converges to $f(\omega)$. Therefore, if $\omega\in\Omega$ there exists k such that if $n>k$, $|f_n(\omega)-f(\omega)|<\frac{1}{m}$ which means $\omega\notin E_{km}$. Note also that $E_{km}\supseteq E_{(k+1)m}$. Since $\mu(E_{1m})<\infty$, Theorem 9.2.4 on Page 246 implies

$$0=\mu(\cap_{k=1}^{\infty}E_{km})=\lim_{k\to\infty}\mu(E_{km}).$$

Let $k(m)$ be chosen such that $\mu(E_{k(m)m})<\varepsilon 2^{-m}$ and let $F=\bigcup_{m=1}^{\infty}E_{k(m)m}$. Then $\mu(F)<\varepsilon$ because

$$\mu(F)\leq\sum_{m=1}^{\infty}\mu\left(E_{k(m)m}\right)<\sum_{m=1}^{\infty}\varepsilon 2^{-m}=\varepsilon$$

Now let $\eta>0$ be given and pick m_0 such that $m_0^{-1}<\eta$. If $\omega\in F^C$, then $\omega\in\bigcap_{m=1}^{\infty}E_{k(m)m}^C$. Hence $\omega\in E_{k(m_0)m_0}^C$ so $d(f(\omega),f_n(\omega))<1/m_0<\eta$ for all $n>k(m_0)$. This holds for all $\omega\in F^C$and so f_n converges uniformly to f on F^C.

Now if $E\neq\emptyset$, consider $\{\mathscr{X}_{E^C}f_n\}_{n=1}^{\infty}$. Then $\mathscr{X}_{E^C}f_n$ is measurable and the sequence converges pointwise to $\mathscr{X}_E f$ everywhere. Therefore, from the first part, there exists a set of measure less than ε,F such that on $F^C,\{\mathscr{X}_{E^C}f_n\}$ converges uniformly to $\mathscr{X}_{E^C}f$. Therefore, on $(E\cup F)^C$, $\{f_n\}$ converges uniformly to f. ■

24.2 The Bochner Integral

24.2.1 Definition and Basic Properties

Definition 24.2.1 *Let $a_k \in X$, a Banach space and let a simple function $\omega \to x(\omega)$ be*

$$x(\omega) = \sum_{k=1}^{n} a_k \mathcal{X}_{E_k}(\omega) \tag{24.3}$$

where for each k, E_k is measurable and $\mu(E_k) < \infty$. Thus this is a measurable finite valued function zero off a set of finite measure. Then define

$$\int_\Omega x(\omega)\, d\mu \equiv \sum_{k=1}^{n} a_k \mu(E_k).$$

Proposition 24.2.2 *Definition 24.2.1 is well defined, the integral is linear on simple functions and*

$$\left\| \int_\Omega x(\omega)\, d\mu \right\| \leq \int_\Omega \|x(\omega)\|\, d\mu$$

whenever x is a simple function.

Proof: It suffices to verify that if $\sum_{k=1}^{n} a_k \mathcal{X}_{E_k}(\omega) = 0$,then $\sum_{k=1}^{n} a_k \mu(E_k) = 0$. Let $f \in X'$. Then

$$f\left(\sum_{k=1}^{n} a_k \mathcal{X}_{E_k}(\omega) \right) = \sum_{k=1}^{n} f(a_k)\, \mathcal{X}_{E_k}(\omega) = 0$$

and, therefore,

$$0 = \int_\Omega \left(\sum_{k=1}^{n} f(a_k)\, \mathcal{X}_{E_k}(\omega) \right) d\mu = \sum_{k=1}^{n} f(a_k)\, \mu(E_k) = f\left(\sum_{k=1}^{n} a_k \mu(E_k) \right).$$

Since $f \in X'$ is arbitrary, and X' separates the points of X, it follows that $\sum_{k=1}^{n} a_k \mu(E_k) = 0$ as hoped. It is now obvious that the integral is linear on simple functions.

As to the triangle inequality, say $x(\omega) = \sum_{k=1}^{n} a_k \mathcal{X}_{E_k}(\omega)$ where the E_k are disjoint. Then from the triangle inequality,

$$\left\| \int_\Omega x(\omega)\, d\mu \right\| = \left\| \sum_{k=1}^{n} a_k \mu(E_k) \right\| \leq \sum_{k=1}^{n} \|a_k\| \mu(E_k) = \int_\Omega \|x(\omega)\|\, d\mu \ \blacksquare$$

Definition 24.2.3 *A strongly measurable function x is Bochner integrable if there exists a sequence of simple functions x_n converging to x pointwise and satisfying*

$$\int_\Omega \|x_n(\omega) - x_m(\omega)\|\, d\mu \to 0 \text{ as } m, n \to \infty. \tag{24.4}$$

If x is Bochner integrable, define

$$\int_\Omega x(\omega)\, d\mu \equiv \lim_{n\to\infty} \int_\Omega x_n(\omega)\, d\mu. \tag{24.5}$$

First it is important to show that this integral is well defined. When this is done, an easier to use condition will be developed. Note that by Lemma 24.1.2, if x is strongly measurable, $\|x\|$ is a measurable real valued function. Thus, it makes sense to consider $\int_\Omega \|x\|\,d\mu$ and also $\int_\Omega \|x-x_n\|\,d\mu$.

Theorem 24.2.4 *The definition of Bochner integrability is well defined. Also, a strongly measurable function x is Bochner integrable if and only if $\int_\Omega \|x\|\,d\mu < \infty$. In this case that the function is Bochner integrable, an approximating sequence of simple functions $\{y_n\}$ exists such that $\|y_n(\omega)\| \le 2\|x(\omega)\|$ for all ω and*

$$\lim_{n\to\infty} \int_\Omega \|y_n(\omega) - x(\omega)\|\,d\mu = 0$$

Proof: $\Rightarrow$First consider the claim about the integral being well defined. Let $\{x_n\}$ be a sequence of simple functions converging pointwise to x and satisfying the conditions given above for x to be Bochner integrable. Then

$$\left| \int_\Omega \|x_n(\omega)\|\,d\mu - \int_\Omega \|x_m(\omega)\|\,d\mu \right| \le \int_\Omega \|x_n - x_m\|\,d\mu$$

which is given to converge to 0 as $n,m \to \infty$ which shows that $\{\int_\Omega \|x_n(\omega)\|\,d\mu\}_{n=1}^\infty$ is a Cauchy sequence. Hence it is bounded and so, by Fatou's lemma,

$$\int_\Omega \|x(\omega)\|\,d\mu \le \liminf_{n\to\infty} \int_\Omega \|x_n(\omega)\|\,d\mu < \infty$$

The limit in 24.5 exists because

$$\left\| \int_\Omega x_n d\mu - \int_\Omega x_m d\mu \right\| = \left\| \int_\Omega (x_n - x_m)\,d\mu \right\| \le \int_\Omega \|x_n - x_m\|\,d\mu$$

and the last term is no more than ε whenever n,m are large enough. From Fatou's lemma, if n is large enough,

$$\int_\Omega \|x_n - x\|\,d\mu < \varepsilon$$

Now if you have another sequence $\{\hat{x}_n\}$ satisfying the condition 24.4 along with pointwise convergence to x,

$$\begin{aligned} \left\| \int_\Omega x_n d\mu - \int_\Omega \hat{x}_n d\mu \right\| &= \left\| \int_\Omega (x_n - \hat{x}_n)\,d\mu \right\| \le \int_\Omega \|x_n - \hat{x}_n\|\,d\mu \\ &\le \int_\Omega \|x_n - x\|\,d\mu + \int_\Omega \|x - \hat{x}_n\|\,d\mu < 2\varepsilon \end{aligned}$$

if n is large enough. Hence convergence of the integrals of the simple functions takes place and these integrals converge to the same thing. Thus the definition is well defined and $\int_\Omega \|x\|\,d\mu < \infty$.

$\Leftarrow$Next suppose $\int_\Omega \|x\|\,d\mu < \infty$ for x strongly measurable. By Lemma 24.1.2, there is a sequence of finite valued measurable functions $\{y_n\}$ with $\|y_n(\omega)\| \le 2\|x(\omega)\|$ and $y_n(\omega) \to x(\omega)$ for each ω. Thus, in fact, y_n is a simple function because it must be zero off a set of finite measure because

$$\int_\Omega \|y_n(\omega)\|\,d\mu < 2\int_\Omega \|x(\omega)\|\,d\mu$$

Then by the dominated convergence theorem for scalar valued functions,

$$\lim_{n\to\infty}\int_\Omega \|y_n - x\|\,d\mu = 0$$

Thus,

$$\int_\Omega \|y_n - y_m\|\,d\mu \le \int_\Omega \|y_n - x\|\,d\mu + \int_\Omega \|x - y_m\|\,d\mu < \varepsilon$$

if m,n are large enough so $\{y_n\}$ is a suitable approximating sequence for x. ■

This is a very nice theorem. It says that all you have to do is verify measurability and absolute integrability just like the case of scalar valued functions. Other things which are totally similar are that the integral is linear, the triangle inequality holds, and you can take a continuous linear functional inside the integral. These things are considered in the following theorem.

Theorem 24.2.5 *The Bochner integral is well defined and if x is Bochner integrable and $f \in X'$,*

$$f\left(\int_\Omega x(\omega)\,d\mu\right) = \int_\Omega f(x(\omega))\,d\mu \tag{24.6}$$

and the triangle inequality is valid,

$$\left\|\int_\Omega x(\omega)\,d\mu\right\| \le \int_\Omega \|x(\omega)\|\,d\mu. \tag{24.7}$$

Also, the Bochner integral is linear. That is, if a,b are scalars and x,y are two Bochner integrable functions, then

$$\int_\Omega (ax(\omega) + by(\omega))\,d\mu = a\int_\Omega x(\omega)\,d\mu + b\int_\Omega y(\omega)\,d\mu \tag{24.8}$$

Proof: Theorem 24.2.4 shows $\int_\Omega \|x(\omega)\|\,d\mu < \infty$ and that the definition of the integral is well defined.

It remains to verify the triangle inequality on Bochner integral functions and the claim about passing a continuous linear functional inside the integral. First of all, consider the triangle inequality. From Lemma 24.1.2, there is a sequence of simple functions $\{y_n\}$ satisfying 24.4 and converging to x pointwise such that also $\|y_n(\omega)\| \le 2\|x(\omega)\|$. Thus,

$$\left\|\int_\Omega x(\omega)\,d\mu\right\| \equiv \lim_{n\to\infty}\left\|\int_\Omega y_n(\omega)\,d\mu\right\| \le \lim_{n\to\infty}\int_\Omega \|y_n(\omega)\|\,d\mu = \int_\Omega \|x(\omega)\|\,d\mu$$

the last step coming from the dominated convergence theorem since $\|y_n(\omega)\| \le 2\|x(\omega)\|$ and $\|y_n(\omega)\| \to \|x(\omega)\|$ for each ω. This shows the triangle inequality.

From Definition 24.2.1 and Theorem 24.2.4 and $\{y_n\}$ being the approximating sequence described there,

$$f\left(\int_\Omega y_n d\mu\right) = \int_\Omega f(y_n)\,d\mu.$$

Thus,

$$f\left(\int_\Omega x d\mu\right) = \lim_{n\to\infty} f\left(\int_\Omega y_n d\mu\right) = \lim_{n\to\infty}\int_\Omega f(y_n)\,d\mu = \int_\Omega f(x)\,d\mu,$$

the last equation holding from the dominated convergence theorem ($|f(y_n)| \leq \|f\| \|y_n\| \leq 2\|f\| \|x\|$). This shows 24.6.

It remains to verify 24.8. Let $f \in X'$. Then from 24.6

$$\begin{aligned} f\left(\int_\Omega (ax(\omega)+by(\omega))\,d\mu\right) &= \int_\Omega (af(x(\omega))+bf(y(\omega)))\,d\mu \\ &= a\int_\Omega f(x(\omega))\,d\mu + b\int_\Omega f(y(\omega))\,d\mu \\ &= f\left(a\int_\Omega x(\omega)\,d\mu + b\int_\Omega y(\omega)\,d\mu\right). \end{aligned}$$

Since X' separates the points of X,it follows

$$\int_\Omega (ax(\omega)+by(\omega))\,d\mu = a\int_\Omega x(\omega)\,d\mu + b\int_\Omega y(\omega)\,d\mu$$

and this proves 24.8. ■

A similar result is the following corollary.

Corollary 24.2.6 *Let an X valued function x be Bochner integrable. Let $L \in \mathscr{L}(X,Y)$ where Y is another Banach space. Then Lx is a Y valued Bochner integrable function and*

$$L\left(\int_\Omega x(\omega)\,d\mu\right) = \int_\Omega Lx(\omega)\,d\mu$$

Proof: From Theorem 24.2.4 there is a sequence of simple functions $\{y_n\}$ having the properties listed in that theorem. These are measurable with finitely many values and are forced to be simple because $\|y_n\| \leq 2\|x\|$. Then consider $\{Ly_n\}$ which converges pointwise to Lx. Since L is continuous and linear,

$$\int_\Omega \|Ly_n - Lx\|_Y\,d\mu \leq \|L\| \int_\Omega \|y_n - x\|_X\,d\mu$$

which converges to 0. This implies

$$\lim_{m,n\to\infty} \int_\Omega \|Ly_n - Ly_m\|\,d\mu = 0$$

and so by definition Lx is Bochner integrable. Also

$$\begin{aligned} \int_\Omega x(\omega)\,d\mu &= \lim_{n\to\infty} \int_\Omega y_n(\omega)\,d\mu \\ \int_\Omega Lx(\omega)\,d\mu &= \lim_{n\to\infty} \int_\Omega Ly_n(\omega)\,d\mu = \lim_{n\to\infty} L\int_\Omega y_n(\omega)\,d\mu \end{aligned}$$

Next,

$$\begin{aligned} &\left\| L\left(\int_\Omega x(\omega)\,d\mu\right) - \int_\Omega Lx(\omega)\,d\mu \right\|_Y \\ &\leq \left\| L\left(\int_\Omega x(\omega)\,d\mu\right) - L\int_\Omega y_n(\omega)\,d\mu \right\|_Y \\ &+ \left\| \int_\Omega Ly_n(\omega)\,d\mu - \int_\Omega Lx(\omega)\,d\mu \right\|_Y < \varepsilon/2 + \varepsilon/2 = \varepsilon \end{aligned}$$

whenever n large enough. ■

24.2.2 Taking a Closed Operator Out of the Integral

Now let X and Y be separable Banach spaces and suppose $A : D(A) \subseteq X \to Y$ be a closed operator. Recall this means that the graph of A,

$$G(A) \equiv \{(x, Ax) : x \in D(A)\}$$

is a closed subset of $X \times Y$ with respect to the product topology obtained from the norm

$$\|(x,y)\| = \max(\|x\|, \|y\|).$$

Thus also $G(A)$ is a separable Banach space with the above norm. You can also consider $D(A)$ as a separable Banach space having the graph norm

$$\|x\|_{D(A)} \equiv \max(\|x\|, \|Ax\|) \tag{24.9}$$

which is isometric to $G(A)$ with the mapping, $\theta x \equiv (x, Ax)$. Recall why this is. It is clear that θ is one to one and onto $G(A)$. Is it continuous? If $x_n \to x$ in $D(A)$, this means that $x_n \to x$ in X and $Ax_n \to y$. Then, since A is closed, it follows that $y = Ax$ so $(x_n, Ax_n) \to (x, Ax)$ in $G(A)$. Hence θ is indeed continuous and onto. Similar reasoning shows that $D(A)$ with this norm is complete. Hence it is a Banach space. Thus θ^{-1} is also continuous. The following lemma is a fundamental result which was proved earlier in the discussion on the Eberlein Smulian theorem in which this was an essential fact to allow the case of a reflexive Banach space which maybe was not separable. See Lemma 21.5.13 for the proof.

Lemma 24.2.7 *A closed subspace of a reflexive Banach space is reflexive.*

Then, with this lemma, one has the following corollary.

Corollary 24.2.8 *Suppose Y is a reflexive Banach space and X is a Banach space such that there exists a continuous one to one mapping, $g : X \to Y$ such that $g(X)$ is a closed subset of Y. Then X is reflexive.*

Proof: By the open mapping theorem, $g(X)$ and X are homeomorphic since g^{-1} must also be continuous. Therefore, since $g(X)$ is reflexive because it is a closed subspace of a reflexive space, it follows X is also reflexive. ■

Lemma 24.2.9 *Suppose V is a reflexive Banach space and that V is a dense subset of W, another Banach space in the topology of W. Then i^*W' is a dense subset of V' where here i is the inclusion map of V into W.*

Proof: First note that i^* is one to one. If $i^*w^* = 0$ for $w^* \in W'$, then this means that for all $v \in V$,

$$i^*w(v) = w^*(v) = 0$$

and since V is dense in W, this shows $w^* = 0$.

Consider the following diagram

$$\begin{array}{ccc} V'' & \overset{i^{**}}{\to} & W'' \\ V' & \overset{i^*}{\leftarrow} & W' \\ V & \overset{i}{\to} & W \end{array}$$

in which i is the inclusion map. Next suppose i^*W' is not dense in V'. Then, using the Hahn Banach theorem, there exists $v^{**} \in V''$ such that $v^{**} \neq 0$ but $v^{**}(i^*W') = 0$. It follows from V being reflexive, that $v^{**} = Jv_0$ where J is the James map from V to $V''$for some $v_0 \in V$. Thus for every $w^* \in W'$,

$$\begin{aligned} 0 &= v^{**}(i^*w^*) \equiv i^{**}v^{**}(w^*) \\ &= i^{**}Jv_0(w^*) = Jv_0(i^*w^*) \\ &\equiv i^*w^*(v_0) = w^*(v_0) \end{aligned}$$

and since W' separates the points of W, it follows $v_0 = 0$ which contradicts $v^{**} \neq 0$. ■

Note that in the proof, only V reflexive was used.

This lemma implies an easy corollary.

Corollary 24.2.10 *Let E and F be reflexive Banach spaces and let A be a closed operator $A : D(A) \subseteq E \to F$. Suppose also that $D(A)$ is dense in E. Then making $D(A)$ into a Banach space by using the above graph norm given in 24.9, it follows that $D(A)$ is a Banach space and i^*E' is a dense subspace of $D(A)'$.*

Proof: First note that $E \times F$ is a reflexive Banach space and $\mathscr{G}(A)$ is a closed subspace of $E \times F$ so it is also a reflexive Banach space. Now $D(A)$ is isometric to $\mathscr{G}(A)$ and so it follows $D(A)$ is a dense subspace of E which is reflexive. Therefore, from Lemma 24.2.9 the conclusion follows. ■

With this preparation, here is another interesting theorem. This one is about taking outside the integral a closed linear operator as opposed to a continuous linear operator.

Theorem 24.2.11 *Let X, Y be separable Banach spaces and let $A : D(A) \subseteq X \to Y$ be a closed operator where $D(A)$ is a dense separable subset of X with respect to the graph norm on $D(A)$ described above[1]. Suppose also that i^*X' is a dense subspace of $D(A)'$ where $D(A)$ is a Banach space having the graph norm described in 24.9. Suppose that $(\Omega, \mathscr{F}, \mu)$ is a measure space and $x : \Omega \to X$ is strongly measurable and it happens that $x(\omega) \in D(A)$ for all $\omega \in \Omega$. Then x is strongly measurable as a mapping into $D(A)$. Also Ax is strongly measurable as a map into Y and if*

$$\int_\Omega \|x(\omega)\| \, d\mu, \int_\Omega \|Ax(\omega)\| \, d\mu < \infty, \tag{24.10}$$

then

$$\int_\Omega x(\omega) \, d\mu \in D(A) \tag{24.11}$$

and

$$A \int_\Omega x(\omega) \, d\mu = \int_\Omega Ax(\omega) \, d\mu. \tag{24.12}$$

Proof: First of all, consider the assertion that x is strongly measurable into $D(A)$. Letting $f \in D(A)'$ be given, there exists a sequence, $\{g_n\} \subseteq i^*X'$ such that $g_n \to f$ in $D(A)'$. Therefore, $\omega \to g_n(x(\omega))$ is measurable by assumption and $g_n(x(\omega)) \to f(x(\omega))$, which shows that $\omega \to f(x(\omega))$ is measurable. By the Pettis theorem, it follows that $\omega \to x(\omega)$ is strongly measurable as a map into $D(A)$.

[1] Note that this follows from the assumed separability of X, Y because the graph is a subset of the separable space $X \times Y$

It follows from Theorem 24.2.4 there exists a sequence of simple functions $\{x_n\}$ of the form

$$x_n(\omega) = \sum_{k=1}^{m_n} a_k^n \mathcal{X}_{E_k^n}(\omega), x_n(\omega) \in D(A),$$

which converges strongly and pointwise to $x(\omega)$ in $D(A)$. Thus

$$x_n(\omega) \to x(\omega), Ax_n(\omega) \to Ax(\omega),$$

which shows $\omega \to Ax(\omega)$ is stongly measurable in Y as claimed.

It remains to verify the assertions about the integral. 24.10 implies x is Bochner integrable as a function having values in $D(A)$ with the norm on $D(A)$ described above. Therefore, by Theorem 24.2.4 there exists a sequence of simple functions $\{y_n\}$ having values in $D(A)$,

$$\lim_{m,n\to\infty} \int_\Omega \|y_n - y_m\|_{D(A)}\, d\mu = 0,$$

$y_n(\omega)$ converging pointwise to $x(\omega)$,

$$\|y_n(\omega)\|_{D(A)} \le 2\|x(\omega)\|_{D(A)}$$

and

$$\lim_{n\to\infty} \int_\Omega \|x(\omega) - y_n(\omega)\|_{D(A)}\, ds = 0.$$

Therefore,

$$\int_\Omega y_n(\omega)\, d\mu \in D(A), \int_\Omega y_n(\omega)\, d\mu \to \int_\Omega x(\omega)\, d\mu \text{ in } X,$$

and since y_n is a simple function and A is linear,

$$A\int_\Omega y_n(\omega)\, d\mu = \int_\Omega Ay_n(\omega)\, d\mu \to \int_\Omega Ax(\omega)\, d\mu \text{ in } Y.$$

It follows, since A is a closed operator, that $\int_\Omega x(\omega)\, d\mu \in D(A)$ and

$$A\int_\Omega x(\omega)\, d\mu = \int_\Omega Ax(\omega)\, d\mu. \blacksquare$$

Here is another version of this theorem which has different hypotheses.

Theorem 24.2.12 *Let X and Y be separable Banach spaces and let $A : D(A) \subseteq X \to Y$ be a closed operator. Also let $(\Omega, \mathscr{F}, \mu)$ be a measure space and let $x : \Omega \to X$ be Bochner integrable such that $x(\omega) \in D(A)$ for all ω. Also suppose Ax is Bochner integrable. Then*

$$\int Ax d\mu = A\int x d\mu$$

and $\int x d\mu \in D(A)$.

Proof: Consider the graph of A,

$$G(A) \equiv \{(x, Ax) : x \in D(A)\} \subseteq X \times Y.$$

Then since A is closed, $G(A)$ is a closed separable Banach space with the norm $\|(x,y)\| \equiv \max(\|x\|, \|y\|)$. Therefore, for $g^* \in G(A)'$, apply the Hahn Banach theorem and obtain

$(x^*,y^*) \in (X \times Y)'$ such that $g^*(x,Ax) = (x^*(x), y^*(Ax))$. Now it follows from the assumptions that $\omega \to (x^*(x(\omega)), y^*(Ax(\omega)))$ is measurable with values in $G(A)$. It is also separably valued because this is true of $G(A)$. By the Pettis theorem, $\omega \to (x(\omega), A(x(\omega)))$ must be strongly measurable. Also $\int \|x(\omega)\| + \|A(x(\omega))\| d\mu < \infty$ by assumption and so there exists a sequence of simple functions having values in $G(A), \{(x_n(\omega), Ax_n(\omega))\}$ which converges to $(x(\omega), A(\omega))$ pointwise such that $\int \|(x_n,Ax_n) - (x,Ax)\| d\mu \to 0$ in $G(A)$. Now for simple functions is it routine to verify that

$$\int (x_n,Ax_n)\, d\mu = \left(\int x_n d\mu, \int Ax_n d\mu \right) = \left(\int x_n d\mu, A \int x_n d\mu \right)$$

Also

$$\begin{aligned} \left\| \int x_n d\mu - \int x d\mu \right\| &\leq \int \|x_n - x\| d\mu \\ &\leq \int \|(x_n,Ax_n) - (x,Ax)\| d\mu \end{aligned}$$

which converges to 0. Also

$$\begin{aligned} \left\| \int Ax_n d\mu - \int Ax d\mu \right\| &= \left\| A \int x_n d\mu - \int Ax d\mu \right\| \\ &\leq \int \|Ax_n - Ax\| d\mu \\ &\leq \int \|(x_n,Ax_n) - (x,Ax)\| d\mu \end{aligned}$$

and this converges to 0. Therefore, $\int x_n d\mu \to \int x d\mu$ and $A \int x_n d\mu \to \int Ax d\mu$. Since each $\int x_n d\mu \in D(A)$, and A is closed, this implies $\int x d\mu \in D(A)$ and $A \int x d\mu = \int Ax d\mu$. ■

24.3 Operator Valued Functions

Consider the case where $A(\omega) \in \mathscr{L}(X,Y)$ for X and Y separable Banach spaces. With the operator norm $\mathscr{L}(X,Y)$ is a Banach space and so if A is strongly measurable, the Bochner integral can be defined as before. However, it is also possible to define the Bochner integral of such operator valued functions for more general situations. In this section, $(\Omega, \mathscr{F}, \mu)$ will be a measure space as usual.

Lemma 24.3.1 *Let $x \in X$ and suppose A is strongly measurable. Then*

$$\omega \to A(\omega) x$$

is strongly measurable as a map into Y.

Proof: Since A is assumed to be strongly measurable, it is the pointwise limit of measurable finite valued functions of the form

$$A_n(\omega) \equiv \sum_{k=1}^{m_n} A_k^n \mathscr{X}_{E_k^n}(\omega)$$

where A_k^n is in $\mathscr{L}(X,Y)$. It follows $A_n(\omega)x \to A(\omega)x$ for each ω and so, since $\omega \to A_n(\omega)x$ is a simple Y valued function, $\omega \to A(\omega)x$ must be strongly measurable. ■

Definition 24.3.2 *Suppose $A(\omega) \in \mathscr{L}(X,Y)$ for each $\omega \in \Omega$ where X, Y are separable Banach spaces. Suppose also that for each $x \in X$,*

$$\omega \to A(\omega)x \text{ is strongly measurable} \tag{24.13}$$

and there exists C such that for each $x \in X$,

$$\int_\Omega \|A(\omega)x\| \, d\mu < C\|x\| \tag{24.14}$$

Then $\int_\Omega A(\omega)\, d\mu \in \mathscr{L}(X,Y)$ is defined by the following formula.

$$\left(\int_\Omega A(\omega)\, d\mu\right)(x) \equiv \int_\Omega A(\omega)x d\mu \tag{24.15}$$

Lemma 24.3.3 *The above definition is well defined. Furthermore, if 24.13 holds then $\omega \to \|A(\omega)\|$ is measurable and if 24.14 holds, then*

$$\left\| \int_\Omega A(\omega)\, d\mu \right\| \leq \int_\Omega \|A(\omega)\| \, d\mu.$$

Proof: It is clear that in case $\omega \to A(\omega)x$ is measurable for all $x \in X$ there exists a unique $\Psi \in \mathscr{L}(X,Y)$ such that

$$\Psi(x) = \int_\Omega A(\omega)x d\mu.$$

This is because $x \to \int_\Omega A(\omega)x d\mu$ is linear and continuous. It is continuous because

$$\left\| \int_\Omega A(\omega)x d\mu \right\| \leq \int_\Omega \|A(\omega)x\| \, d\mu \leq \int_\Omega \|A(\omega)\| \, d\mu \, \|x\|$$

Thus $\Psi = \int_\Omega A(\omega)\, d\mu$ and the definition is well defined.

Now consider the assertion about $\omega \to \|A(\omega)\|$. Let $D' \subseteq B'$ the closed unit ball in Y' be such that D' is countable and

$$\|y\| = \sup_{y^* \in D'} |y^*(y)|.$$

This is from Lemma 24.1.7. Recall X is separable. Also let D be a countable dense subset of B, the unit ball of X. Then

$$\begin{aligned} \{\omega : \|A(\omega)\| > \alpha\} &= \left\{\omega : \sup_{x \in D} \|A(\omega)x\| > \alpha\right\} \\ &= \cup_{x \in D} \{\omega : \|A(\omega)x\| > \alpha\} \\ &= \cup_{x \in D} \left(\cup_{y^* \in D'} \{|y^*(A(\omega)x)| > \alpha\}\right) \end{aligned}$$

and this is measurable because $\omega \to A(\omega)x$ is strongly, hence weakly measurable.

Now suppose 24.14 holds. Then for all x,

$$\int_\Omega \|A(\omega)x\| \, d\mu < C\|x\|.$$

It follows that for $\|x\| \leq 1$,

$$\left\| \left(\int_\Omega A(\omega)\, d\mu \right)(x) \right\| = \left\| \int_\Omega A(\omega)\, x\, d\mu \right\| \leq \int_\Omega \|A(\omega)\, x\|\, d\mu \leq \int_\Omega \|A(\omega)\|\, d\mu$$

and so

$$\left\| \int_\Omega A(\omega)\, d\mu \right\| \leq \int_\Omega \|A(\omega)\|\, d\mu. \blacksquare$$

Now it is interesting to consider the case where $A(\omega) \in \mathscr{L}(H,H)$ where $\omega \to A(\omega)x$ is strongly measurable and $A(\omega)$ is compact and self adjoint. Recall the Kuratowski measurable selection theorem, Theorem 9.15.8 on Page 278 listed here for convenience.

Theorem 24.3.4 *Let E be a compact metric space and let $(\Omega, \mathscr{F})$ be a measure space. Suppose $\psi : E \times \Omega \to \mathbb{R}$ has the property that $x \to \psi(x,\omega)$ is continuous and $\omega \to \psi(x,\omega)$ is measurable. Then there exists a measurable function, f having values in E such that*

$$\psi(f(\omega),\omega) = \sup_{x \in E} \psi(x,\omega).$$

Furthermore, $\omega \to \psi(f(\omega),\omega)$ is measurable.

24.3.1 Review of Hilbert Schmidt Theorem

This section is a review of earlier material and is presented a little differently. I think it does not hurt to repeat some things relative to Hilbert space. I will give a proof of the Hilbert Schmidt theorem which will generalize to a result about measurable operators. It will be a little different then the earlier proof. Recall the following.

Definition 24.3.5 *Define $v \otimes u \in \mathscr{L}(H,H)$ by*

$$v \otimes u(x) = (x,u)\, v.$$

$A \in \mathscr{L}(H,H)$ is a compact operator if whenever $\{x_k\}$ is a bounded sequence, there exists a convergent subsequence of $\{Ax_k\}$. Equivalently, A maps bounded sets to sets whose closures are compact or to use other terminology, A maps bounded sets to sets which are precompact.

Next is a convenient description of compact operators on a Hilbert space.

Lemma 24.3.6 *Let H be a Hilbert space and suppose $A \in \mathscr{L}(H,H)$ is a compact operator. Then*

1. *A is a compact operator if and only if whenever if $x_n \to x$ weakly in H, it follows that $Ax_n \to Ax$ strongly in H.*
2. *For $u, v \in H$, $v \otimes u : H \to H$ is a compact operator.*
3. *Let B be the closed unit ball in H. If A is self adjoint and compact, then if $x_n \to x$ weakly on B, it follows that $(Ax_n, x_n) \to (Ax, x)$ so $x \to |(Ax,x)|$ achieves its maximum value on B.*
4. *The function, $v \otimes u$ is compact and the operator $u \otimes u$ is self adjoint.*

Proof: Consider $\Rightarrow$ of 1. Suppose then that $x_n \to x$ weakly. Since $\{x_n\}$ is weakly bounded, it follows from the uniform boundedness principle that $\{\|x_n\|\}$ is bounded. Let $x_n \in \hat{B}$ for $\hat{B}$ some closed ball. If Ax_n fails to converge to Ax, then there is $\varepsilon > 0$ and a subsequence still denoted as $\{x_n\}$ such that $x_n \to x$ weakly but $\|Ax_n - Ax\| \geq \varepsilon > 0$. Then $A\left(\hat{B}\right)$ is precompact because A is compact so there is a further subsequence, still denoted by $\{x_n\}$ such that Ax_n converges to some $y \in H$. Therefore,

$$\begin{aligned}(y,w) &= \lim_{n\to\infty}(Ax_n,w) = \lim_{n\to\infty}(x_n,A^*w)\\ &= (x,A^*w) = (Ax,w)\end{aligned}$$

which shows $Ax = y$ since w is arbitrary. However, this is a contradiction to $\|Ax_n - Ax\| \geq \varepsilon > 0$.

Consider $\Leftarrow$ of 1. Why is A compact if it satisfies the property that it takes weakly convergent sequences to strongly convergent ones? If A is not compact, then there exists $\hat{B}$ a bounded set such that $A\left(\hat{B}\right)$ is not precompact. Thus, there exists a sequence $\{Ax_n\}_{n=1}^{\infty} \subseteq A\left(\hat{B}\right)$ which has no convergent subsequence where $x_n \in \hat{B}$ the bounded set. However, there is a subsequence $\{x_n\} \in \hat{B}$ which converges weakly to some $x \in H$ because of weak compactness. Hence $Ax_n \to Ax$ by assumption and so this is a contradiction to there being no convergent subsequence of $\{Ax_n\}_{n=1}^{\infty}$.

Next consider 2. Letting $\{x_n\}$ be a bounded sequence,

$$v \otimes u(x_n) = (x_n,u)\,v.$$

There exists a weakly convergent subsequence of $\{x_n\}$ say $\left\{x_{n_k}\right\}$ converging weakly to $x \in H$. Therefore,

$$\left\|v \otimes u\left(x_{n_k}\right) - v \otimes u(x)\right\| = \left\|\left(x_{n_k},u\right) - (x,u)\right\| \|v\|$$

which converges to 0. Thus $v \otimes u$ is compact as claimed. It takes bounded sets to precompact sets.

Next consider 3. To verify the assertion about $x \to (Ax,x)$, let $x_n \to x$ weakly. Since A is compact, $Ax_n \to Ax$ by part 1. Then, since A is self adjoint,

$$\begin{aligned}&|(Ax_n,x_n) - (Ax,x)|\\ \leq\ &|(Ax_n,x_n) - (Ax,x_n)| + |(Ax,x_n) - (Ax,x)|\\ \leq\ &|(Ax_n,x_n) - (Ax,x_n)| + |(Ax_n,x) - (Ax,x)|\\ \leq\ &\|Ax_n - Ax\|\,\|x_n\| + \|Ax_n - Ax\|\,\|x\| \leq 2\,\|Ax_n - Ax\|\end{aligned}$$

which converges to 0. Now let $\{x_n\}$ be a maximizing sequence for $|(Ax,x)|$ for $x \in B$ and let $\lambda \equiv \sup\{|(Ax,x)| : x \in B\}$. There is a subsequence still denoted as $\{x_n\}$ which converges weakly to some $x \in B$ by weak compactness. Hence $|(Ax,x)| = \lim_{n\to\infty}|(Ax_n,x_n)| = \lambda$.

Next consider 4. It only remains to verify that $u \otimes u$ is self adjoint. This follows from the definition.

$$\begin{aligned}((u \otimes u)x,y) &\equiv (u(x,u),y) = (x,u)(u,y)\\ (x,(u \otimes u)y) &\equiv (x,u(y,u)) = (u,y)(x,u),\end{aligned}$$

the same thing. ■

Observation 24.3.7 *Note that if A is any self adjoint operator,*

$$\overline{(Ax,x)} = (x,Ax) = (Ax,x).$$

so (Ax,x) is real valued.

From Lemma 24.3.6, the maximum of $|(Ax,x)|$ exists on the closed unit ball B.

Lemma 24.3.8 *Let $A \in \mathscr{L}(H,H)$ and suppose it is self adjoint and compact. Let B denote the closed unit ball in H. Let $e \in B$ be such that*

$$|(Ae,e)| = \max_{x\in B} |(Ax,x)|.$$

Then letting $\lambda = (Ae,e)$, it follows $Ae = \lambda e$. You can always assume $\|e\| = 1$.

Proof: From the above observation, (Ax,x) is always real and since A is compact, $|(Ax,x)|$ achieves a maximum at e. It remains to verify e is an eigenvector. If $|(Ae,e)| = 0$ for all $e \in B$, then A is a self adjoint nonnegative ($(Ax,x) \geq 0$) operator and so by Cauchy Schwarz inequality,

$$(Ae,x) \leq (Ax,x)^{1/2} (Ae,e)^{1/2} = 0$$

and so $Ae = 0$ for all e. Assume then that A is not 0. You can always make $|(Ae,e)|$ at least as large by replacing e with $e/\|e\|$. Thus, there is no loss of generality in letting $\|e\| = 1$ in every case.

Suppose $\lambda = (Ae,e) \geq 0$ where $|(Ae,e)| = \max_{x\in B} |(Ax,x)|$. Thus

$$((\lambda I - A)e,e) = \lambda \|e\|^2 - \lambda = 0$$

Then it is easy to verify that $\lambda I - A$ is a nonnegative ($((\lambda I - A)x,x) \geq 0$ for all x.) and self adjoint operator. To see this, note that

$$((\lambda I - A)x,x) = \|x\|^2 \left((\lambda I - A)\frac{x}{\|x\|}, \frac{x}{\|x\|} \right) = \|x\|^2 \lambda - \|x\|^2 \left(A\frac{x}{\|x\|}, \frac{x}{\|x\|} \right) \geq 0$$

Therefore, the Cauchy Schwarz inequality can be applied to write

$$((\lambda I - A)e,x) \leq ((\lambda I - A)e,e)^{1/2} ((\lambda I - A)x,x)^{1/2} = 0$$

Since this is true for all x it follows $Ae = \lambda e$. Just pick $x = (\lambda I - A)e$.

Next suppose $\max_{x\in B} |(Ax,x)| = -(Ae,e)$. Let $-\lambda = (-Ae,e)$ and the previous result can be applied to $-A$ and $-\lambda$. Thus $-\lambda e = -Ae$ and so $Ae = \lambda e$. ■

With these lemmas here is a major theorem, the Hilbert Schmidt theorem. I think this proof is a little slicker than the more standard proof given earlier.

Theorem 24.3.9 *Let $A \in \mathscr{L}(H,H)$ be a compact self adjoint operator on a Hilbert space. Then there exist real numbers $\{\lambda_k\}_{k=1}^{\infty}$ and vectors $\{e_k\}_{k=1}^{\infty}$ such that*

$$\|e_k\| = 1,$$

$$(e_k,e_j)_H = 0 \text{ if } k \neq j,$$

$$Ae_k = \lambda_k e_k,$$

$$|\lambda_n| \geq |\lambda_{n+1}| \text{ for all } n,$$

$$\lim_{n\to\infty} \lambda_n = 0,$$

$$\lim_{n\to\infty} \left\| A - \sum_{k=1}^{n} \lambda_k (e_k \otimes e_k) \right\|_{\mathscr{L}(H,H)} = 0. \tag{24.16}$$

Proof: This is done by considering a sequence of compact self adjoint operators, $A, A_1, A_2, \cdots$. Here is how these are defined. Using Lemma 24.3.8 let e_1, λ_1 be given by that lemma such that

$$|(Ae_1, e_1)| = \max_{x\in B} |(Ax, x)|, \ \lambda_1 = (Ae_1, e_1) \Rightarrow Ae_1 = \lambda_1 e_1$$

Then by that lemma, $Ae_1 = \lambda_1 e_1$ and $\|e_1\| = 1$. Now define $A_1 = A - \lambda_1 e_1 \otimes e_1$. This is compact and self adjoint by Lemma 24.3.6. Thus, one could repeat the argument.

If A_n has been obtained, use Lemma 24.3.8 to obtain e_{n+1} and λ_{n+1} such that

$$|(A_n e_{n+1}, e_{n+1})| = \max_{x\in B} |(A_n x, x)|, \ \lambda_{n+1} = (A_n e_{n+1}, e_{n+1}).$$

By that lemma again, $A_n e_{n+1} = \lambda_{n+1} e_{n+1}$ and $\|e_{n+1}\| = 1$. Then

$$A_{n+1} \equiv A_n - \lambda_{n+1} e_{n+1} \otimes e_{n+1}$$

Thus iterating this,

$$A_n = A - \sum_{k=1}^{n} \lambda_k e_k \otimes e_k. \tag{24.17}$$

Assume for $j, k \leq n, (e_k, e_j) = \delta_{jk}$. Then the new vector e_{n+1} will be orthogonal to the earlier ones. This is the next claim.

Claim 1: If $k < n+1$ then $(e_{n+1}, e_k) = 0$. Also $Ae_k = \lambda_k e_k$ for all k and from the construction, $A_n e_{n+1} = \lambda_{n+1} e_{n+1}$.

Proof of claim: From the above,

$$\lambda_{n+1} e_{n+1} = A_n e_{n+1} = Ae_{n+1} - \sum_{k=1}^{n} \lambda_k (e_{n+1}, e_k) e_k.$$

From the above and induction hypothesis that $(e_k, e_j) = \delta_{jk}$ for $j, k \leq n$,

$$\begin{aligned} \lambda_{n+1} (e_{n+1}, e_j) &= (Ae_{n+1}, e_j) - \sum_{k=1}^{n} \lambda_k (e_{n+1}, e_k) (e_k, e_j) \\ &= (e_{n+1}, Ae_j) - \sum_{k=1}^{n} \lambda_k (e_{n+1}, e_k) (e_k, e_j) \\ &= \lambda_j (e_{n+1}, e_j) - \lambda_j (e_{n+1}, e_j) = 0. \end{aligned}$$

To verify the second part of this claim,

$$\lambda_{n+1} e_{n+1} = A_n e_{n+1} = Ae_{n+1} - \sum_{k=1}^{n} \lambda_k e_k (e_{n+1}, e_k) = Ae_{n+1}$$

This proves the claim.

Claim 2: $|\lambda_n| \geq |\lambda_{n+1}|$.

Proof of claim: From 24.17 and the definition of A_n and $e_k \otimes e_k$,

$$\begin{aligned}(A_{n-1}e_{n+1}, e_{n+1}) &= \left(\left(A - \sum_{k=1}^{n-1} \lambda_k e_k \otimes e_k\right) e_{n+1}, e_{n+1}\right) \\ &= (Ae_{n+1}, e_{n+1}) = (A_n e_{n+1}, e_{n+1})\end{aligned}$$

Thus,

$$\begin{aligned}\lambda_{n+1} &= (A_n e_{n+1}, e_{n+1}) \\ &= (A_{n-1}e_{n+1}, e_{n+1}) - \lambda_n |(e_n, e_{n+1})|^2 \\ &= (A_{n-1}e_{n+1}, e_{n+1})\end{aligned}$$

By the previous claim. Therefore,

$$|\lambda_{n+1}| = |(A_{n-1}e_{n+1}, e_{n+1})| \leq |(A_{n-1}e_n, e_n)| = |\lambda_n|$$

by the definition of $|\lambda_n|$. (e_n makes $|(A_{n-1}x, x)|$ as large as possible.)

Claim 3: $\lim_{n\to\infty} \lambda_n = 0$.

Proof of claim: If for some $n, \lambda_n = 0$, then $\lambda_k = 0$ for all $k > n$ by claim 2. Thus, for some n,

$$A = \sum_{k=1}^{n} \lambda_k e_k \otimes e_k$$

Assume then that $\lambda_k \neq 0$ for any k. Then if $\lim_{k\to\infty} |\lambda_k| = \varepsilon > 0$, one contradicts, $\|e_k\| = 1$ for all k because

$$\begin{aligned}\|Ae_n - Ae_m\|^2 &= \|\lambda_n e_n - \lambda_m e_m\|^2 \\ &= \lambda_n^2 + \lambda_m^2 \geq 2\varepsilon^2\end{aligned}$$

which shows there is no Cauchy subsequence of $\{Ae_n\}_{n=1}^{\infty}$, which contradicts the compactness of A. This proves the claim.

Claim 4: $\|A_n\| \to 0$

Proof of claim: Let $x, y \in B$

$$\begin{aligned}|\lambda_{n+1}| &\geq \left|\left(A_n \frac{x+y}{2}, \frac{x+y}{2}\right)\right| \\ &= \left|\frac{1}{4}(A_n x, x) + \frac{1}{4}(A_n y, y) + \frac{1}{2}(A_n x, y)\right| \\ &\geq \frac{1}{2}|(A_n x, y)| - \frac{1}{4}|(A_n x, x) + (A_n y, y)| \\ &\geq \frac{1}{2}|(A_n x, y)| - \frac{1}{4}(|(A_n x, x)| + |(A_n y, y)|) \\ &\geq \frac{1}{2}|(A_n x, y)| - \frac{1}{2}|\lambda_{n+1}|\end{aligned}$$

and so

$$3|\lambda_{n+1}| \geq |(A_n x, y)|.$$

It follows $\|A_n\| \leq 3|\lambda_{n+1}|$. By 24.17 this proves 24.16 and completes the proof. ■

24.3.2 Measurable Compact Operators

Here the operators will be of the form $A(\omega)$ where $\omega \in \Omega$ and $\omega \to A(\omega)x$ is strongly measurable and $A(\omega)$ is a compact operator in $\mathscr{L}(H,H)$.

Theorem 24.3.10 *Let $A(\omega) \in \mathscr{L}(H,H)$ be a compact self adjoint operator and H is a separable Hilbert space such that $\omega \to A(\omega)x$ is strongly measurable. Then there exist real numbers $\{\lambda_k(\omega)\}_{k=1}^{\infty}$ and vectors $\{e_k(\omega)\}_{k=1}^{\infty}$ such that*

$$\|e_k(\omega)\| = 1$$

$$(e_k(\omega), e_j(\omega))_H = 0 \text{ if } k \neq j,$$

$$A(\omega)e_k(\omega) = \lambda_k(\omega)e_k(\omega),$$

$$|\lambda_n(\omega)| \geq |\lambda_{n+1}(\omega)| \text{ for all } n,$$

$$\lim_{n\to\infty} \lambda_n(\omega) = 0,$$

$$\lim_{n\to\infty} \left\| A(\omega) - \sum_{k=1}^{n} \lambda_k(\omega)(e_k(\omega) \otimes e_k(\omega)) \right\|_{\mathscr{L}(H,H)} = 0.$$

The function $\omega \to \lambda_j(\omega)$ is measurable and $\omega \to e_j(\omega)$ is strongly measurable.

Proof: It is simply a repeat of the above proof of the Hilbert Schmidt theorem except at every step when the e_k and λ_k are defined, you use the Kuratowski measurable selection theorem, Theorem 24.3.4 on Page 687 to obtain $\lambda_k(\omega)$ is measurable and that $\omega \to e_k(\omega)$ is also measurable. This follows because the closed unit ball in a separable Hilbert space is a compact metric space.

When you consider $\max_{x\in B} |(A_n(\omega)x,x)|$, let $\psi(x,\omega) = |(A_n(\omega)x,x)|$. Then ψ is continuous in x by Lemma 24.3.6 on Page 687 and it is measurable in ω by assumption. Therefore, by the Kuratowski theorem, $e_k(\omega)$ is measurable in the sense that inverse images of weakly open sets in B are measurable. However, by Lemma 24.1.12 on Page 674 this is the same as weakly measurable. Since H is separable, this implies $\omega \to e_k(\omega)$ is also strongly measurable. The measurability of λ_k and e_k is the only new thing here and so this completes the proof. ■

24.4 Fubini's Theorem for Bochner Integrals

Now suppose $(\Omega_1, \mathscr{F}, \mu)$ and $(\Omega_2, \mathscr{S}, \lambda)$ are two σ finite measure spaces. Recall the notion of product measure. There was a σ algebra, denoted by $\mathscr{F} \times \mathscr{S}$ which is the smallest σ algebra containing the elementary sets, (finite disjoint unions of measurable rectangles) and a measure, denoted by $\mu \times \lambda$ defined on this σ algebra such that for $E \in \mathscr{F} \times \mathscr{S}$,

$$s_1 \to \lambda(E_{s_1}), \ (E_{s_1} \equiv \{s_2 : (s_1,s_2) \in E\})$$

is μ measurable and

$$s_2 \to \mu(E_{s_2}), \ (E_{s_2} \equiv \{s_1 : (s_1,s_2) \in E\})$$

is λ measurable. In terms of nonnegative functions which are $\mathscr{F}\times\mathscr{S}$ measurable,

$$\begin{aligned} s_1 &\to f(s_1,s_2) \text{ is } \mu \text{ measurable,}\\ s_2 &\to f(s_1,s_2) \text{ is } \lambda \text{ measurable,}\\ s_1 &\to \int_{\Omega_2} f(s_1,s_2)\,d\lambda \text{ is } \mu \text{ measurable,}\\ s_2 &\to \int_{\Omega_1} f(s_1,s_2)\,d\mu \text{ is } \lambda \text{ measurable,} \end{aligned}$$

and the conclusion of Fubini's theorem holds.

$$\begin{aligned} \int_{\Omega_1\times\Omega_2} f\,d(\mu\times\lambda) &= \int_{\Omega_1}\int_{\Omega_2} f(s_1,s_2)\,d\lambda\,d\mu\\ &= \int_{\Omega_2}\int_{\Omega_1} f(s_1,s_2)\,d\mu\,d\lambda. \end{aligned}$$

The following theorem is the version of Fubini's theorem valid for Bochner integrable functions.

Theorem 24.4.1 *Let $f:\Omega_1\times\Omega_2\to X$ be strongly measurable with respect to $\mu\times\lambda$ and suppose*

$$\int_{\Omega_1\times\Omega_2} ||f(s_1,s_2)||\,d(\mu\times\lambda)<\infty. \tag{24.18}$$

Then there exist a set of μ measure zero, N and a set of λ measure zero, M such that the following formula holds with all integrals making sense.

$$\begin{aligned} \int_{\Omega_1\times\Omega_2} f(s_1,s_2)\,d(\mu\times\lambda) &= \int_{\Omega_1}\int_{\Omega_2} f(s_1,s_2)\,\mathscr{X}_{N^C}(s_1)\,d\lambda\,d\mu\\ &= \int_{\Omega_2}\int_{\Omega_1} f(s_1,s_2)\,\mathscr{X}_{M^C}(s_2)\,d\mu\,d\lambda. \end{aligned}$$

Proof: First note that from 24.18 and the usual Fubini theorem for nonnegative valued functions,

$$\int_{\Omega_1\times\Omega_2} ||f(s_1,s_2)||\,d(\mu\times\lambda)=\int_{\Omega_1}\int_{\Omega_2} ||f(s_1,s_2)||\,d\lambda\,d\mu$$

and so

$$\int_{\Omega_2} \|f(s_1,s_2)\|\,d\lambda<\infty \tag{24.19}$$

for μ a.e. s_1. Say for all $s_1\notin N$ where $\mu(N)=0$.

Let $\phi\in X'$. Then $\phi\circ f$ is $\mathscr{F}\times\mathscr{S}$ measurable and

$$\begin{aligned} &\int_{\Omega_1\times\Omega_2} |\phi\circ f(s_1,s_2)|\,d(\mu\times\lambda)\\ \le\ &\int_{\Omega_1\times\Omega_2} \|\phi\|\,\|f(s_1,s_2)\|\,d(\mu\times\lambda)<\infty \end{aligned}$$

and so from the usual Fubini theorem for complex valued functions,

$$\int_{\Omega_1\times\Omega_2} \phi\circ f(s_1,s_2)\,d(\mu\times\lambda)=\int_{\Omega_1}\int_{\Omega_2} \phi\circ f(s_1,s_2)\,d\lambda\,d\mu. \tag{24.20}$$

Now also if you fix s_2, it follows from the definition of strongly measurable and the properties of product measure mentioned above that

$$s_1 \to f(s_1, s_2)$$

is strongly measurable. Also, by 24.19

$$\int_{\Omega_2} \|f(s_1, s_2)\| \, d\lambda < \infty$$

for $s_1 \notin N$. Therefore, by Theorem 24.2.4 $s_2 \to f(s_1, s_2)\, \mathcal{X}_{N^C}(s_1)$ is Bochner integrable. By 24.20 and 24.6

$$\begin{aligned} & \int_{\Omega_1 \times \Omega_2} \phi \circ f(s_1, s_2) \, d(\mu \times \lambda) \\ = & \int_{\Omega_1} \int_{\Omega_2} \phi \circ f(s_1, s_2) \, d\lambda d\mu \\ = & \int_{\Omega_1} \int_{\Omega_2} \phi(f(s_1, s_2)\, \mathcal{X}_{N^C}(s_1)) \, d\lambda d\mu \\ = & \int_{\Omega_1} \phi\left(\int_{\Omega_2} f(s_1, s_2)\, \mathcal{X}_{N^C}(s_1) \, d\lambda \right) d\mu. \end{aligned} \tag{24.21}$$

Each iterated integral makes sense and

$$\begin{aligned} s_1 \to & \int_{\Omega_2} \phi(f(s_1, s_2)\, \mathcal{X}_{N^C}(s_1)) \, d\lambda \\ = & \phi\left(\int_{\Omega_2} f(s_1, s_2)\, \mathcal{X}_{N^C}(s_1) \, d\lambda \right) \end{aligned} \tag{24.22}$$

is μ measurable because

$$\begin{aligned} (s_1, s_2) \to & \phi(f(s_1, s_2)\, \mathcal{X}_{N^C}(s_1)) \\ = & \phi(f(s_1, s_2))\, \mathcal{X}_{N^C}(s_1) \end{aligned}$$

is product measurable. Now consider the function,

$$s_1 \to \int_{\Omega_2} f(s_1, s_2)\, \mathcal{X}_{N^C}(s_1) \, d\lambda. \tag{24.23}$$

I want to show this is also Bochner integrable with respect to μ so I can factor out ϕ once again. It's measurability follows from the Pettis theorem and the above observation 24.22. Also,

$$\begin{aligned} & \int_{\Omega_1} \left\| \int_{\Omega_2} f(s_1, s_2)\, \mathcal{X}_{N^C}(s_1) \, d\lambda \right\| d\mu \\ \leq & \int_{\Omega_1} \int_{\Omega_2} \|f(s_1, s_2)\| \, d\lambda d\mu \\ = & \int_{\Omega_1 \times \Omega_2} \|f(s_1, s_2)\| \, d(\mu \times \lambda) < \infty. \end{aligned}$$

Therefore, the function in 24.23 is indeed Bochner integrable and so in 24.21 the ϕ can be taken outside the last integral. Thus,

$$\begin{aligned}
&\phi\left(\int_{\Omega_1\times\Omega_2} f(s_1,s_2)\,d(\mu\times\lambda)\right)\\
=&\int_{\Omega_1\times\Omega_2} \phi\circ f(s_1,s_2)\,d(\mu\times\lambda)\\
=&\int_{\Omega_1}\int_{\Omega_2} \phi\circ f(s_1,s_2)\,d\lambda d\mu\\
=&\int_{\Omega_1}\phi\left(\int_{\Omega_2} f(s_1,s_2)\,\mathcal{X}_{N^C}(s_1)\,d\lambda\right)d\mu\\
=&\phi\left(\int_{\Omega_1}\int_{\Omega_2} f(s_1,s_2)\,\mathcal{X}_{N^C}(s_1)\,d\lambda d\mu\right).
\end{aligned}$$

Since X' separates the points,

$$\int_{\Omega_1\times\Omega_2} f(s_1,s_2)\,d(\mu\times\lambda)=\int_{\Omega_1}\int_{\Omega_2} f(s_1,s_2)\,\mathcal{X}_{N^C}(s_1)\,d\lambda d\mu.$$

The other formula follows from similar reasoning. ■

24.5 The Spaces $L^p(\Omega;X)$

Recall that x is Bochner when it is strongly measurable and $\int_\Omega \|x(s)\|\,d\mu<\infty$. It is natural to generalize to $\int_\Omega \|x(s)\|^p\,d\mu<\infty$.

Definition 24.5.1 *$x\in L^p(\Omega;X)$ for $p\in[1,\infty)$ if x is strongly measurable and*

$$\int_\Omega \|x(s)\|^p\,d\mu<\infty$$

Also

$$\|x\|_{L^p(\Omega;X)}\equiv\|x\|_p\equiv\left(\int_\Omega \|x(s)\|^p\,d\mu\right)^{1/p}. \tag{24.24}$$

As in the case of scalar valued functions, two functions in $L^p(\Omega;X)$ are considered equal if they are equal a.e. With this convention, and using the same arguments found in the presentation of scalar valued functions it is clear that $L^p(\Omega;X)$ is a normed linear space with the norm given by 24.24. In fact, $L^p(\Omega;X)$ is a Banach space. This is the main contribution of the next theorem.

Lemma 24.5.2 *If x_n is a Cauchy sequence in $L^p(\Omega;X)$ satisfying*

$$\sum_{n=1}^{\infty}\|x_{n+1}-x_n\|_p<\infty,$$

then there exists $x\in L^p(\Omega;X)$ such that $x_n(s)\to x(s)$ a.e. and

$$\|x-x_n\|_p\to 0.$$

Proof: Let $g_N(s) \equiv \sum_{n=1}^{N} \|x_{n+1}(s) - x_n(s)\|_X$. Then by the triangle inequality,

$$\begin{aligned}\left(\int_\Omega g_N(s)^p \, d\mu\right)^{1/p} &\leq \sum_{n=1}^{N} \left(\int_\Omega \|x_{n+1}(s) - x_n(s)\|^p \, d\mu\right)^{1/p} \\ &\leq \sum_{n=1}^{\infty} \|x_{n+1} - x_n\|_p < \infty.\end{aligned}$$

Let

$$g(s) = \lim_{N\to\infty} g_N(s) = \sum_{n=1}^{\infty} \|x_{n+1}(s) - x_n(s)\|_X.$$

By the monotone convergence theorem,

$$\left(\int_\Omega g(s)^p \, d\mu\right)^{1/p} = \lim_{N\to\infty} \left(\int_\Omega g_N(s)^p \, d\mu\right)^{1/p} < \infty.$$

Therefore, there exists a measurable set of measure 0 called E, such that for $s \notin E$, $g(s) < \infty$. Hence, for $s \notin E$, $\lim_{N\to\infty} x_{N+1}(s)$ exists because

$$x_{N+1}(s) = x_{N+1}(s) - x_1(s) + x_1(s) = \sum_{n=1}^{N} (x_{n+1}(s) - x_n(s)) + x_1(s).$$

Thus, if $N > M$, and s is a point where $g(s) < \infty$,

$$\begin{aligned}\|x_{N+1}(s) - x_{M+1}(s)\|_X &\leq \sum_{n=M+1}^{N} \|x_{n+1}(s) - x_n(s)\|_X \\ &\leq \sum_{n=M+1}^{\infty} \|x_{n+1}(s) - x_n(s)\|_X\end{aligned}$$

which shows that $\{x_{N+1}(s)\}_{N=1}^{\infty}$ is a Cauchy sequence for each $s \notin E$. Now let

$$x(s) \equiv \begin{cases} \lim_{N\to\infty} x_N(s) \text{ if } s \notin E, \\ 0 \text{ if } s \in E. \end{cases}$$

Theorem 24.1.10 shows that x is strongly measurable. By Fatou's lemma,

$$\int_\Omega \|x(s) - x_N(s)\|^p \, d\mu \leq \liminf_{M\to\infty} \int_\Omega \|x_M(s) - x_N(s)\|^p \, d\mu.$$

But if N and M are large enough with $M > N$,

$$\left(\int_\Omega \|x_M(s) - x_N(s)\|^p \, d\mu\right)^{1/p} \leq \sum_{n=N}^{M} \|x_{n+1} - x_n\|_p \leq \sum_{n=N}^{\infty} \|x_{n+1} - x_n\|_p < \varepsilon$$

and this shows, since ε is arbitrary, that

$$\lim_{N\to\infty} \int_\Omega \|x(s) - x_N(s)\|^p \, d\mu = 0.$$

It remains to show $x \in L^p(\Omega; X)$. This follows from the above and the triangle inequality. Thus, for N large enough,

$$\left(\int_\Omega \|x(s)\|^p \, d\mu\right)^{1/p} \leq \left(\int_\Omega \|x_N(s)\|^p \, d\mu\right)^{1/p}$$

$$+\left(\int_\Omega \|x(s)-x_N(s)\|^p\,d\mu\right)^{1/p} \le \left(\int_\Omega \|x_N(s)\|^p\,d\mu\right)^{1/p}+\varepsilon<\infty. \blacksquare$$

Theorem 24.5.3 *$L^p(\Omega;X)$ is complete. Also every Cauchy sequence has a subsequence which converges pointwise.*

Proof: If $\{x_n\}$ is Cauchy in $L^p(\Omega;X)$, extract a subsequence $\{x_{n_k}\}$ satisfying

$$\left\|x_{n_{k+1}}-x_{n_k}\right\|_p \le 2^{-k}$$

and apply Lemma 24.5.2. The pointwise convergence of this subsequence was established in the proof of this lemma. This proves the theorem because if a subsequence of a Cauchy sequence converges, then the Cauchy sequence must also converge. ■

Observation 24.5.4 *If the measure space is Lebesgue measure then you have continuity of translation in $L^p(\mathbb{R}^n;X)$ in the usual way. More generally, for μ a Radon measure on Ω a locally compact Hausdorff space, $C_c(\Omega;X)$ is dense in $L^p(\Omega;X)$. Here $C_c(\Omega;X)$ is the space of continuous X valued functions which have compact support in Ω. The proof of this little observation follows immediately from approximating with simple functions and then applying the appropriate considerations to the simple functions.*

Clearly Fatou's lemma and the monotone convergence theorem make no sense for functions with values in a Banach space but the dominated convergence theorem holds in this setting.

Theorem 24.5.5 *If x is strongly measurable and $x_n(s)\to x(s)$ a.e. (for s off a set of measure zero) with*

$$\|x_n(s)\| \le g(s) \text{ a.e.}$$

where $\int_\Omega g\,d\mu<\infty$, then x is Bochner integrable and

$$\int_\Omega x(s)\,d\mu = \lim_{n\to\infty}\int_\Omega x_n(s)\,d\mu.$$

Proof: The measurability of x follows from Theorem 24.1.10 if convergence happens for each s. Otherwise, x is measurable by assumption. Then $\|x_n(s)-x(s)\| \le 2g(s)$ a.e. so, from Fatou's lemma,

$$\begin{aligned}\int_\Omega 2g(s)\,d\mu &\le \liminf_{n\to\infty}\int_\Omega (2g(s)-\|x_n(s)-x(s)\|)\,d\mu\\ &= \int_\Omega 2g(s)\,d\mu - \limsup_{n\to\infty}\int_\Omega \|x_n(s)-x(s)\|\,d\mu\end{aligned}$$

and so,

$$\limsup_{n\to\infty}\int_\Omega \|x_n(s)-x(s)\|\,d\mu \le 0$$

Also, from Fatou's lemma again,

$$\int_\Omega \|x(s)\|\,d\mu \le \liminf_{n\to\infty}\int_\Omega \|x_n(s)\|\,d\mu < \int_\Omega g(s)\,d\mu<\infty$$

so $x\in L^1$. Then by the triangle inequality,

$$\limsup_{n\to\infty}\left\|\int_\Omega x(s)\,d\mu-\int_\Omega x_n(s)\,d\mu\right\| \le \limsup_{n\to\infty}\int_\Omega \|x_n(s)-x(s)\|\,d\mu = 0 \blacksquare$$

One can also give a version of the Vitali convergence theorem.

Definition 24.5.6 *Let $\mathscr{A} \subseteq L^1(\Omega;X)$. Then $\mathscr{A}$ is said to be uniformly integrable if for every $\varepsilon > 0$ there exists $\delta > 0$ such that whenever $\mu(E) < \delta$, it follows*

$$\int_E \|f\|_X \, d\mu < \varepsilon$$

for all $f \in \mathscr{A}$. It is bounded if

$$\sup_{f \in \mathscr{A}} \int_\Omega \|f\|_X \, d\mu < \infty.$$

Theorem 24.5.7 *Let $(\Omega, \mathscr{F}, \mu)$ be a finite measure space and let X be a separable Banach space. Let $\{f_n\} \subseteq L^1(\Omega;X)$ be uniformly integrable and bounded such that $f_n(\omega) \to f(\omega)$ for each $\omega \in \Omega$. Then $f \in L^1(\Omega;X)$ and*

$$\lim_{n\to\infty} \int_\Omega \|f_n - f\|_X \, d\mu = 0.$$

Proof: Let $\varepsilon > 0$ be given. Then by uniform integrability there exists $\delta > 0$ such that if $\mu(E) < \delta$ then

$$\int_E \|f_n\| \, d\mu < \varepsilon/3.$$

By Fatou's lemma the same inequality holds for f. Fatou's lemma shows $f \in L^1(\Omega;X)$, f being measurable because of Theorem 9.1.2.

By Egoroff's theorem, Theorem 24.1.18, there exists a set of measure less than δ, E such that the convergence of $\{f_n\}$ to f is uniform off E. Therefore,

$$\begin{aligned} \int_\Omega \|f - f_n\| \, d\mu &\leq \int_E (\|f\|_X + \|f_n\|_X) \, d\mu + \int_{E^C} \|f - f_n\|_X \, d\mu \\ &< \frac{2\varepsilon}{3} + \int_{E^C} \frac{\varepsilon}{(\mu(\Omega)+1)\,3} d\mu < \varepsilon \end{aligned}$$

if n is large enough. ■

Note that a convenient way to achieve uniform integrability is to say $\{f_n\}$ is bounded in $L^p(\Omega;X)$ for some $p > 1$. This follows from Holder's inequality.

$$\int_E \|f_n\| \, d\mu \leq \left(\int_E d\mu \right)^{1/p'} \left(\int_\Omega \|f_n\|^p \, d\mu \right)^{1/p} \leq C\mu(E)^{1/p'}.$$

The following theorem is interesting.

Theorem 24.5.8 *Let $1 \leq p < \infty$ and let $p < r \leq \infty$. Then $L^r([0,T];X)$ is a Borel subset of $L^p([0,T];X)$. Letting $C([0,T];X)$ denote the continuous functions having values in X, $C([0,T];X)$ is also a Borel subset of $L^p([0,T];X)$. Here the measure is ordinary one dimensional Lebesgue measure on $[0,T]$.*

Proof: First consider the claim about $L^r([0,T];X)$. Let

$$B_M \equiv \left\{ x \in L^p([0,T];X) : \|x\|_{L^r([0,T];X)} \leq M \right\}.$$

Then B_M is a closed subset of $L^p([0,T];X)$. Here is why. If $\{x_n\}$ is a sequence of elements of B_M and $x_n \to x$ in $L^p([0,T];X)$, then passing to a subsequence, still denoted by x_n, it can be assumed $x_n(s) \to x(s)$ a.e. Hence Fatou's lemma can be applied to conclude

$$\int_0^T \|x(s)\|^r ds \leq \lim\inf_{n\to\infty} \int_0^T \|x_n(s)\|^r ds \leq M^r < \infty.$$

Now $\cup_{M=1}^{\infty} B_M = L^r([0,T];X)$. Note this did not depend on the measure space used. It would have been equally valid on any measure space.

Consider now $C([0,T];X)$. The norm on this space is the usual norm, $\|\cdot\|_\infty$. The argument above shows $\|\cdot\|_\infty$ is a Borel measurable function on $L^p([0,T];X)$. This is because $B_M \equiv \{x \in L^p([0,T];X) : \|x\|_\infty \leq M\}$ is a closed, hence Borel subset of $L^p([0,T];X)$. Now let $\theta \in \mathscr{L}(L^p([0,T];X), L^p(\mathbb{R};X))$ such that $\theta(x(t)) = x(t)$ for all $t \in [0,T]$ and also $\theta \in \mathscr{L}(C([0,T];X), BC(\mathbb{R};X))$ where $BC(\mathbb{R};X)$ denotes the bounded continuous functions with a norm given by $\|x\| \equiv \sup_{t\in\mathbb{R}} \|x(t)\|$, and θx has compact support.

For example, you could define

$$\widetilde{x}(t) \equiv \begin{cases} x(t) \text{ if } t \in [0,T] \\ x(2T-t) \text{ if } t \in [T,2T] \\ x(-t) \text{ if } t \in [-T,0] \\ 0 \text{ if } t \notin [-T,2T] \end{cases}$$

and let $\Phi \in C_c^\infty(-T,2T)$ such that $\Phi(t) = 1$ for $t \in [0,T]$. Then you could let

$$\theta x(t) \equiv \Phi(t)\widetilde{x}(t).$$

Then let $\{\phi_n\}$ be a mollifier and define

$$\psi_n x(t) \equiv \phi_n * \theta x(t).$$

It follows $\psi_n x$ is uniformly continuous because

$$\begin{aligned} & \left\| \psi_n x(t) - \psi_n x(t') \right\|_X \\ \leq & \int_{\mathbb{R}} \left| \phi_n(t'-s) - \phi_n(t-s) \right| \|\theta x(s)\|_X ds \\ \leq & C\|x\|_p \left(\int_{\mathbb{R}} \left| \phi_n(t'-s) - \phi_n(t-s) \right|^{p'} ds \right)^{1/p'} \end{aligned}$$

Also for $x \in C([0,T];X)$, it follows from usual mollifier arguments that

$$\|\psi_n x - x\|_{L^\infty([0,T];X)} \to 0.$$

Here is why. For $t \in [0,T]$,

$$\begin{aligned} \|\psi_n x(t) - x(t)\|_X & \leq \int_{\mathbb{R}} \phi_n(s) \|\theta x(t-s) - \theta x(t)\| ds \\ & \leq C_\theta \int_{-1/n}^{1/n} \phi_n(s) ds \varepsilon = C_\theta \varepsilon \end{aligned}$$

provided n is large enough due to the compact support and consequent uniform continuity of θx.

If $||\psi_n x - x||_{L^\infty([0,T];X)} \to 0$, then $\{\psi_n x\}$ is a Cauchy sequence in $C([0,T];X)$ and this requires that x equals a continuous function a.e. Thus $C([0,T];X)$ consists exactly of those functions, x of $L^p([0,T];X)$ such that $\|\psi_n x - x\|_\infty \to 0$. It follows

$$C([0,T];X) =$$

$$\cap_{n=1}^\infty \cup_{m=1}^\infty \cap_{k=m}^\infty \left\{ x \in L^p([0,T];X) : \|\psi_k x - x\|_\infty \leq \frac{1}{n} \right\}. \tag{24.25}$$

It only remains to show

$$S \equiv \left\{ x \in L^p([0,T];X) : \|\psi_k x - x\|_\infty \leq \alpha \right\}$$

is a Borel set. Suppose then that $x_n \in S$ and $x_n \to x$ in $L^p([0,T];X)$. Then there exists a subsequence, still denoted by n such that $x_n \to x$ pointwise a.e. as well as in L^p. There exists a set of measure 0 such that for all n, and t not in this set,

$$\begin{aligned} \|\psi_k x_n(t) - x_n(t)\| &\equiv \left\| \int_{-1/k}^{1/k} \phi_k(s) (\theta x_n(t-s))\, ds - x_n(t) \right\| \leq \alpha \\ x_n(t) &\to x(t). \end{aligned}$$

Then

$$\begin{aligned} & \|\psi_k x_n(t) - x_n(t) - (\psi_k x(t) - x(t))\| \\ \leq\ & \|x_n(t) - x(t)\|_X + \left\| \int_{-1/k}^{1/k} \phi_k(s) (\theta x_n(t-s) - \theta x(t-s))\, ds \right\| \\ \leq\ & \|x_n(t) - x(t)\|_X + C_{k,\theta} \|x_n - x\|_{L^p(0,T;X)} \end{aligned}$$

which converges to 0 as $n \to \infty$. It follows that for a.e. t,

$$\|\psi_k x(t) - x(t)\| \leq \alpha.$$

Thus S is closed and so the set in 24.25 is a Borel set. ■

As in the scalar case, the following lemma holds in this more general context.

Lemma 24.5.9 *Let (Ω,μ) be a regular measure space where Ω is a locally compact Hausdorff space or more simply a metric space with closed balls compact. Then $C_c(\Omega;X)$ the space of continuous functions having compact support and values in X is dense in $L^p(0,T;X)$ for all $p \in [0,\infty)$. For any measure space $(\Omega,\mathcal{F},\mu)$, the simple functions are dense in $L^p(0,T;X)$.*

Proof: First, the simple functions are dense in $L^p(0,T;X)$. Let $f \in L^p(0,T;X)$ and let $\{x_n\}$ denote a sequence of simple functions which converge to f pointwise which also have the property that

$$\|x_n(s)\| \leq 2\|f(s)\|$$

Then

$$\int_\Omega \|x_n(s) - f(s)\|^p\, d\mu \to 0$$

from the dominated convergence theorem. Therefore, the simple functions are indeed dense in $L^p(0,T;X)$.

Next suppose (Ω,μ) is a regular measure space. If $x(s)\equiv\sum_i a_i\mathscr{X}_{E_i}(s)$ is a simple function, then by regularity, there exist compact sets K_i and open sets, V_i such that $K_i\subseteq E_i\subseteq V_i$ and $\mu(V_i\setminus K_i)^{1/p}<\varepsilon/\sum_i\|a_i\|$. Let $K_i\prec h_i\prec V_i$. Then consider

$$\sum_i a_i h_i\in C_c(\Omega).$$

By the triangle inequality,

$$\begin{aligned}
&\left(\int_\Omega\left\|\sum_i a_i h_i(s)-a_i\mathscr{X}_{E_i}(s)\right\|^p d\mu\right)^{1/p}\\
\leq\ &\sum_i\left(\int_\Omega\|a_i(h_i(s)-\mathscr{X}_{E_i}(s))\|^p d\mu\right)^{1/p}\\
\leq\ &\sum_i\left(\int_\Omega\|a_i\|^p|h_i(s)-\mathscr{X}_{E_i}(s)|^p d\mu\right)^{1/p}\leq\sum_i\|a_i\|\left(\int_{V_i\setminus K_i}d\mu\right)^{1/p}\\
\leq\ &\sum_i\|a_i\|\mu(V_i\setminus K_i)^{1/p}<\varepsilon
\end{aligned}$$

This and the first part of the lemma shows that $\overline{C_c(\Omega;X)}=L^p(\Omega;X)$. ■

24.6 Measurable Representatives

In this section consider the special case where $X=L^1(B,\nu)$ where $(B,\mathscr{F},\nu)$ is a measure space and $x\in L^1(\Omega;X)$. Thus for each $s\in\Omega$, $x(s)\in L^1(B,\nu)$. In general, the map

$$(s,t)\to x(s)(t)$$

will not be product measurable, but one can obtain a measurable representative. This is important because it allows the use of Fubini's theorem on the measurable representative.

By Theorem 24.2.4, there exists a sequence of simple functions, $\{x_n\}$, of the form

$$x_n(s)=\sum_{k=1}^m a_k\mathscr{X}_{E_k}(s)\tag{24.26}$$

where $a_k\in L^1(B,\nu)$ which satisfy the conditions of Definition 24.2.3 and

$$\|x_n-x_m\|_{L^1(\Omega,L^1(B))}\to 0\text{ as }m,n\to\infty\tag{24.27}$$

For such a simple function, you can assume the E_k are disjoint and then

$$\begin{aligned}
\|x_n\|_{L^1(\Omega,L^1(B))}&=\sum_{k=1}^m\|a_k\|_{L^1(B)}\mu(E_k)=\sum_{k=1}^m\int_B|a_k|\,d\nu\mu(E_k)\\
&=\int_\Omega\int_B|a_k(t)|\,d\nu(t)\,\mathscr{X}_{E_k}(s)\,d\mu(s)\\
&=\int_\Omega\int_B|x_n|\,d\nu d\mu
\end{aligned}$$

Also, each x_n is product measurable. Thus from 24.27,

$$\|x_n-x_m\|_{L^1(\Omega,L^1(B))}=\int_\Omega\int_B|x_n-x_m|\,d\nu d\mu$$

which shows that $\{x_n\}$ is a Cauchy sequence in $L^1(\Omega\times B,\mu\times\lambda)$. Then there exists $y\in L^1(\Omega\times B,\mu\times\lambda)$ and a subsequence still called $\{x_n\}$ such that

$$\begin{aligned}\lim_{n\to\infty}\int_\Omega\int_B |x_n-y|\,d\nu d\mu &= \lim_{n\to\infty}\int_\Omega \|x_n-y\|_{L^1(B)}\,d\mu\\ &= \|x_n-y\|_{L^1(\Omega,L^1(B))}=0.\end{aligned}$$

Now consider 24.27. Since $\lim_{m\to\infty}x_m(s)=x(s)$ in $L^1(B)$, it follows from Fatou's lemma that

$$\|x_n-x\|_{L^1(\Omega,L^1(B))}\le \liminf_{m\to\infty}\|x_n-x_m\|_{L^1(\Omega,L^1(B))}<\varepsilon$$

for all n large enough. Hence

$$\lim_{n\to\infty}\|x_n-x\|_{L^1(\Omega,L^1(B))}=0$$

and so

$$x(s)=y(s) \text{ in } L^1(B) \ \mu \text{ a.e. } s$$

In particular, for a.e. s, it follows that

$$x(s)(t)=y(s,t) \text{ for a.e. } t.$$

Now $\int_\Omega x(s)\,d\mu\in X=L^1(B,\nu)$ so it makes sense to ask for $(\int_\Omega x(s)\,d\mu)(t)$, at least μ a.e. t. To find what this is, note

$$\left\|\int_\Omega x_n(s)\,d\mu-\int_\Omega x(s)\,d\mu\right\|_X\le\int_\Omega\|x_n(s)-x(s)\|_X\,d\mu.$$

Therefore, since the right side converges to 0,

$$\lim_{n\to\infty}\left\|\int_\Omega x_n(s)\,d\mu-\int_\Omega x(s)\,d\mu\right\|_X=$$

$$\lim_{n\to\infty}\int_B\left|\left(\int_\Omega x_n(s)\,d\mu\right)(t)-\left(\int_\Omega x(s)\,d\mu\right)(t)\right|d\nu=0.$$

But

$$\left(\int_\Omega x_n(s)\,d\mu\right)(t)=\int_\Omega x_n(s,t)\,d\mu \text{ a.e. } t.$$

Therefore

$$\lim_{n\to\infty}\int_B\left|\int_\Omega x_n(s,t)\,d\mu-\left(\int_\Omega x(s)\,d\mu\right)(t)\right|d\nu=0. \tag{24.28}$$

Also, since $x_n\to y$ in $L^1(\Omega\times B)$,

$$0=\lim_{n\to\infty}\int_B\int_\Omega|x_n(s,t)-y(s,t)|\,d\mu d\nu\ge$$

$$\lim_{n\to\infty}\int_B\left|\int_\Omega x_n(s,t)\,d\mu-\int_\Omega y(s,t)\,d\mu\right|d\nu. \tag{24.29}$$

From 24.28 and 24.29

$$\int_\Omega y(s,t)\,d\mu=\left(\int_\Omega x(s)\,d\mu\right)(t) \text{ a.e. } t.$$

Thus the following theorem is obtained.

Theorem 24.6.1 *Let $X = L^1(B)$ where $(B, \mathscr{F}, \nu)$ is a σ finite measure space and let $x \in L^1(\Omega; X)$. Then there exists a measurable representative, $y \in L^1(\Omega \times B)$, such that*

$$x(s) = y(s,\cdot) \text{ a.e. } s \text{ in } \Omega, \text{ the equation in } L^1(B),$$

and

$$\int_\Omega y(s,t)\, d\mu = \left(\int_\Omega x(s)\, d\mu \right)(t) \text{ a.e. } t.$$

24.7 Vector Measures

There is also a concept of vector measures.

Definition 24.7.1 *Let $(\Omega, \mathscr{S})$ be a set and a σ algebra of subsets of Ω. A mapping*

$$F : \mathscr{S} \to X$$

is said to be a vector measure if

$$F\left(\cup_{i=1}^\infty E_i\right) = \sum_{i=1}^\infty F(E_i)$$

whenever $\{E_i\}_{i=1}^\infty$ is a sequence of disjoint elements of $\mathscr{S}$. For F a vector measure,

$$|F|(A) \equiv \sup\{ \sum_{F \in \pi(A)} \|\mu(F)\| : \pi(A) \text{ is a partition of } A\}.$$

This is the same definition that was given in the case where F would have values in $\mathbb{C}$, the only difference being the fact that now F has values in a general Banach space X as the vector space of values of the vector measure. Recall that a partition of A is a finite set, $\{F_1, \cdots, F_m\} \subseteq \mathscr{S}$ such that $\cup_{i=1}^m F_i = A$. The same theorem about $|F|$ proved in the case of complex valued measures holds in this context with the same proof. For completeness, it is included here.

Theorem 24.7.2 *If $|F|(\Omega) < \infty$, then $|F|$ is a measure on $\mathscr{S}$.*

Proof: Let E_1 and E_2 be sets of $\mathscr{S}$ such that $E_1 \cap E_2 = \emptyset$ and let $\{A_1^i, \cdots, A_{n_i}^i\} = \pi(E_i)$, a partition of E_i which is chosen such that

$$|F|(E_i) - \varepsilon < \sum_{j=1}^{n_i} \|F(A_j^i)\| \quad i = 1, 2.$$

Consider the sets which are contained in either of $\pi(E_1)$ or $\pi(E_2)$, it follows this collection of sets is a partition of $E_1 \cup E_2$ which is denoted here by $\pi(E_1 \cup E_2)$. Then by the above inequality and the definition of total variation,

$$|F|(E_1 \cup E_2) \geq \sum_{F \in \pi(E_1 \cup E_2)} \|F(F)\| > |F|(E_1) + |F|(E_2) - 2\varepsilon,$$

which shows that since $\varepsilon > 0$ was arbitrary,

$$|F|(E_1 \cup E_2) \geq |F|(E_1) + |F|(E_2). \tag{24.30}$$

Let $\{E_j\}_{j=1}^{\infty}$ be a sequence of disjoint sets of $\mathscr{S}$ and let $E_\infty = \cup_{j=1}^{\infty} E_j$. Then by the definition of total variation there exists a partition of E_∞, $\pi(E_\infty) = \{A_1, \cdots, A_n\}$ such that

$$|F|(E_\infty) - \varepsilon < \sum_{i=1}^{n} \|F(A_i)\|.$$

Also,

$$A_i = \cup_{j=1}^{\infty} A_i \cap E_j, \text{ so } F(A_j) = \sum_{j=1}^{\infty} F(A_i \cap E_j)$$

and so by the triangle inequality, $\|F(A_i)\| \leq \sum_{j=1}^{\infty} \|F(A_i \cap E_j)\|$. Therefore, by the above,

$$|F|(E_\infty) - \varepsilon < \sum_{i=1}^{n} \overbrace{\sum_{j=1}^{\infty} \|F(A_i \cap E_j)\|}^{\geq \|F(A_i)\|} = \sum_{j=1}^{\infty} \sum_{i=1}^{n} \|F(A_i \cap E_j)\| \leq \sum_{j=1}^{\infty} |F|(E_j)$$

because $\{A_i \cap E_j\}_{i=1}^{n}$ is a partition of E_j.

Since $\varepsilon > 0$ is arbitrary, this shows

$$|F|(\cup_{j=1}^{\infty} E_j) \leq \sum_{j=1}^{\infty} |F|(E_j).$$

Also, 24.30 implies that whenever the E_i are disjoint, $|F|(\cup_{j=1}^{n} E_j) \geq \sum_{j=1}^{n} |F|(E_j)$. Therefore,

$$\sum_{j=1}^{\infty} |F|(E_j) \geq |F|(\cup_{j=1}^{\infty} E_j) \geq |F|(\cup_{j=1}^{n} E_j) \geq \sum_{j=1}^{n} |F|(E_j).$$

Since n is arbitrary,

$$|F|(\cup_{j=1}^{\infty} E_j) = \sum_{j=1}^{\infty} |F|(E_j)$$

which shows that $|F|$ is a measure as claimed. ■

Definition 24.7.3 *A Banach space is said to have the Radon Nikodym property if whenever*

$$(\Omega, \mathscr{S}, \mu) \text{ is a finite measure space}$$

$$F : \mathscr{S} \to X \text{ is a vector measure with } |F|(\Omega) < \infty$$

$$F \ll \mu$$

then one may conclude there exists $g \in L^1(\Omega; X)$ such that

$$F(E) = \int_E g(s)\, d\mu$$

for all $E \in \mathscr{S}$.

Some Banach spaces have the Radon Nikodym property and some don't. No attempt is made to give a complete answer to the question of which Banach spaces have this property, but the next theorem gives examples of many spaces which do. This next lemma was used earlier. I am presenting it again.

Lemma 24.7.4 *Suppose ν is a complex measure defined on $\mathcal{S}$ a σ algebra where $(\Omega,\mathcal{S})$ is a measurable space, and let μ be a measure on $\mathcal{S}$ with $|\nu(E)| \leq r\mu(E)$ and suppose there is $h \in L^1(\Omega,\mu)$ such that for all $E \in \mathcal{S}$,*

$$\nu(E) = \int_E h d\mu,,$$

Then $|h| \leq r$ a.e.

Proof: Let $B(p,\delta) \subseteq \mathbb{C} \setminus \overline{B(0,r)}$ and let $E \equiv h^{-1}(B(p,\delta))$. If $\mu(E) > 0$. Then

$$\left| \frac{1}{\mu(E)} \int_E h d\mu - p \right| \leq \frac{1}{\mu(E)} \int_E |h(\omega) - p| \, d\mu < \delta$$

Thus, $\left| \frac{\nu(E)}{\mu(E)} - p \right| < \delta$ and so $|\nu(E) - p\mu(E)| < \delta\mu(E)$ which implies

$$|\nu(E)| \geq (|p| - \delta)\mu(E) > r\mu(E) \geq |\nu(E)|$$

which contradicts the assumption. Hence $h^{-1}(B(p,\delta))$ is a set of μ measure zero for all such balls contained in $\mathbb{C} \setminus \overline{B(0,r)}$ and so, since countably many of these balls cover $\mathbb{C} \setminus \overline{B(0,r)}$, it follows that $\mu\left(h^{-1}\left(\mathbb{C} \setminus \overline{B(0,r)}\right)\right) = 0$ and so $|h(\omega)| \leq r$ for a.e. ω. ■

Theorem 24.7.5 *Suppose X' is a separable dual space. Then X' has the Radon Nikodym property.*

Proof: By Theorem 24.1.16, X is separable. Let D be a countable dense subset of X. Let $F \ll \mu, \mu$ a finite measure and F a vector measure and let $|F|(\Omega) < \infty$. Pick $x \in X$ and consider the map

$$E \to F(E)(x)$$

for $E \in \mathcal{S}$. This defines a complex measure which is absolutely continuous with respect to $|F|$. Therefore, by the earlier Radon Nikodym theorem, there exists $f_x \in L^1(\Omega,|F|)$ such that

$$F(E)(x) = \int_E f_x(s) \, d|F|. \tag{24.31}$$

Also, by definition $\|F(E)\| \leq |F|(E)$ so $|F(E)(x)| \leq |F|(E)\|x\|$. By Lemma 24.7.4, $|f_x(s)| \leq \|x\|$ for $|F|$ a.e. s. Let $\tilde{D}$ consist of all finite linear combinations of the form $\sum_{i=1}^m a_i x_i$ where a_i is a rational point of $\mathbb{F}$ and $x_i \in D$. For each of these countably many vectors, there is an exceptional set of measure zero off which $|f_x(s)| \leq \|x\|$. Let N be the union of all of them and define $f_x(s) \equiv 0$ if $s \notin N$. Then since $F(E)$ is in X', it is linear and so for $\sum_{i=1}^m a_i x_i \in \tilde{D}$,

$$\begin{aligned} \int_E f_{\sum_{i=1}^m a_i x_i}(s) \, d|F| &= F(E)\left(\sum_{i=1}^m a_i x_i\right) = \sum_{i=1}^m a_i F(E)(x_i) \\ &= \int_E \sum_{i=1}^m a_i f_{x_i}(s) \, d|F| \end{aligned}$$

and so by uniqueness in the Radon Nikodym theorem,

$$f_{\sum_{i=1}^m a_i x_i}(s) = \sum_{i=1}^m a_i f_{x_i}(s) \ \ |F| \text{ a.e.}$$

and so, we can regard this as holding for all $s \notin N$. Also, if $x \in \tilde{D}, |f_x(s)| \leq \|x\|$. Now for $x, y \in \tilde{D}$,

$$\left|f_x(s) - f_y(s)\right| = \left|f_{x-y}(s)\right| \leq \|x-y\|$$

and so, by density of $\tilde{D}$, we can define

$$h_x(s) \equiv \lim_{n\to\infty} f_{x_n}(s) \text{ where } x_n \to x, x_n \in \tilde{D}$$

For $s \in N$, all functions equal 0. Thus for all x, $|h_x(s)| \leq \|x\|$. The dominated convergence theorem and continuity of $F(E)$ implies that for $x_n \to x$, with $x_n \in \tilde{D}$,

$$\int_E h_x(s)\, d|F| = \lim_{n\to\infty} \int_E f_{x_n}(s)\, d|F| = \lim_{n\to\infty} F(E)(x_n) = F(E)(x). \tag{24.32}$$

It follows from the density of $\tilde{D}$ that for all $x, y \in X, s \notin N$, and $a, b \in \mathbb{F}$, let $x_n \to x, y_n \to y, a_n \to a, b_n \to b$, with $x_n, y_n \in \tilde{D}$ and $a_n, b_n \in \mathbb{Q}$ or $\mathbb{Q} + i\mathbb{Q}$ in case $\mathbb{F} = \mathbb{C}$. Then

$$h_{ax+by}(s) = \lim_{n\to\infty} f_{a_n x_n + b_n y_n}(s) = \lim_{n\to\infty} a_n f_{x_n}(s) + b_n f_{y_n}(s) \equiv a h_x(s) + b h_y(s), \tag{24.33}$$

Let $\theta(s)$ be given by $\theta(s)(x) = h_x(s)$ if $s \notin N$ and let $\theta(s) = 0$ if $s \in N$. By 24.33 it follows that $\theta(s) \in X'$ for each s. Also

$$\theta(s)(x) = h_x(s) \in L^1(\Omega)$$

so $\theta(\cdot)$ is weak $*$ measurable. Since X' is separable, Theorem 24.1.15 implies that θ is strongly measurable. Furthermore, by 24.33,

$$\|\theta(s)\| \equiv \sup_{\|x\|\leq 1} |\theta(s)(x)| \leq \sup_{\|x\|\leq 1} |h_x(s)| \leq 1.$$

Therefore, $\int_\Omega \|\theta(s)\|\, d|F| < \infty$ so $\theta \in L^1(\Omega; X')$. Thus, if $E \in \mathscr{S}$,

$$\int_E h_x(s)\, d|F| = \int_E \theta(s)(x)\, d|F| = \left(\int_E \theta(s)\, d|F|\right)(x). \tag{24.34}$$

From 24.32 and 24.34, $\left(\int_E \theta(s)\, d|F|\right)(x) = F(E)(x)$ for all $x \in X$ and therefore,

$$\int_E \theta(s)\, d|F| = F(E).$$

Finally, since $F \ll \mu, |F| \ll \mu$ also and so there exists $k \in L^1(\Omega)$ such that

$$|F|(E) = \int_E k(s)\, d\mu$$

for all $E \in \mathscr{S}$, by the scalar Radon Nikodym Theorem. It follows

$$F(E) = \int_E \theta(s)\, d|F| = \int_E \theta(s) k(s)\, d\mu.$$

Letting $g(s) = \theta(s) k(s)$, this has proved the theorem. ■

Since each reflexive Banach spaces is a dual space, the following corollary holds.

Corollary 24.7.6 *Any separable reflexive Banach space has the Radon Nikodym property.*

It is not necessary to assume separability in the above corollary. For the proof of a more general result, consult *Vector Measures* by Diestal and Uhl, [13].

24.8 The Riesz Representation Theorem

The Riesz representation theorem for the spaces $L^p(\Omega;X)$ holds under certain conditions. The proof follows the proofs given earlier for scalar valued functions.

Definition 24.8.1 *If X and Y are two Banach spaces, X is isometric to Y if there exists $\theta \in \mathscr{L}(X,Y)$ such that*

$$\|\theta x\|_Y = \|x\|_X .$$

This will be written as $X \cong Y$. The map θ is called an isometry.

The next theorem says that $L^{p'}(\Omega;X')$ is always isometric to a subspace of $(L^p(\Omega;X))'$ for any Banach space, X.

Theorem 24.8.2 *Let X be any Banach space and let $(\Omega,\mathscr{S},\mu)$ be a measure space. Then for $p>1$, $L^{p'}(\Omega;X')$ is isometric to a subspace of $(L^p(\Omega;X))'$. Also, for $g \in L^{p'}(\Omega;X')$,*

$$\sup_{\|f\|_p \le 1} \left| \int_\Omega g(\omega)(f(\omega))\, d\mu \right| = \|g\|_{p'} .$$

If $p=1$ and $p'=\infty$, this is still true assuming $\mu(\Omega)<\infty$.

Proof: First observe that for $f \in L^p(\Omega;X)$ and $g \in L^{p'}(\Omega;X')$,

$$\omega \to g(\omega)(f(\omega))$$

is a function in $L^1(\Omega)$. (To obtain measurability, write f as a limit of simple functions. Holder's inequality then yields the function is in $L^1(\Omega)$.) Define

$$\theta : L^{p'}(\Omega;X') \to (L^p(\Omega;X))'$$

by

$$\theta g(f) \equiv \int_\Omega g(\omega)(f(\omega))\, d\mu .$$

Holder's inequality implies

$$\|\theta g\| \le \|g\|_{p'} \tag{24.35}$$

and it is also clear that θ is linear. Next it is required to show $\|\theta g\| = \|g\|_{p'}$. To begin with, always assume $p>1$.

This will first be verified for simple functions. Assume $\|g\|_{p'} \neq 0$ since if not, there is nothing to show. Let

$$g(\omega) = \sum_{i=1}^m c_i^* \mathscr{X}_{E_i}(\omega),\ g \in L^{p'}(\Omega;X'), c_i^* \neq 0$$

where $0 \neq c_i^* \in X'$, the E_i are disjoint. Let $d_i \in X$ be such that $\|d_i\|_X = 1$ and

$$c_i^*(d_i) \ge \|c_i^*\| - \varepsilon$$

Then let

$$f(\omega) \equiv \frac{1}{\|g\|^{p'-1}_{L^{p'}(\Omega;X')}} \sum_{i=1}^m d_i \|c_i^*\|^{p'-1} \mathscr{X}_{E_i}(\omega)$$

Then since $p'-1=p'/p$,

$$\int_\Omega \|f\|^p\,d\mu = \frac{1}{\|g\|^{p'}_{L^{p'}(\Omega;X')}}\int_\Omega \sum_{i=1}^m \|c_i^*\|^{p'}\,\mathcal{X}_{E_i}(\omega)\,d\mu = \frac{\|g\|^{p'}_{L^{p'}(\Omega;X')}}{\|g\|^{p'}_{L^{p'}(\Omega;X')}} = 1$$

Also $\int_\Omega g(\omega)(f(\omega)) =$

$$\begin{aligned}
&\frac{1}{\|g\|^{p'-1}_{L^{p'}(\Omega;X')}}\int_\Omega \sum_{i=1}^m c_i^*(d_i)\,\|c_i^*\|^{p'-1}\,\mathcal{X}_{E_i}(\omega)\,d\mu\\
\geq\ &\frac{1}{\|g\|^{p'-1}_{L^{p'}(\Omega;X')}}\int_\Omega \sum_{i=1}^m (\|c_i^*\|-\varepsilon)\,\|c_i^*\|^{p'-1}\,\mathcal{X}_{E_i}(\omega)\,d\mu\\
=\ &\frac{1}{\|g\|^{p'-1}_{L^{p'}(\Omega;X')}}\sum_{i=1}^m \|c_i^*\|^{p'}\,\mu(E_i) - \frac{1}{\|g\|^{p'-1}_{L^{p'}(\Omega;X')}}\varepsilon\sum_{i=1}^m \|c_i^*\|^{p'-1}\,\mu(E_i)\\
=\ &\|g\|_{L^{p'}(\Omega;X')}-\varepsilon
\end{aligned}$$

Therefore, $\|g\|_{L^{p'}(\Omega;X')} \geq \|\theta(g)\| \geq |\int_\Omega g(\omega)(f(\omega))| \geq \|g\|_{L^{p'}(\Omega;X')}-\varepsilon$ and since ε is arbitrary, it follows that $\|g\|_{L^{p'}(\Omega;X')} = \|\theta(g)\|$ whenever g is a simple function.

In general, let $g\in L^{p'}(\Omega;X')$ and let g_n be a sequence of simple functions converging to g in $L^{p'}(\Omega;X')$. Such a sequence exists by Lemma 24.1.2. Let $g_n(\omega)\to g(\omega), \|g_n(\omega)\| \leq 2\|g(\omega)\|$. Then each g_n is in $L^{p'}(\Omega;X')$ and by the dominated convergence theorem they converge to g in $L^{p'}(\Omega;X')$. Then for $\|\cdot\|$ the norm in $(L^p(\Omega;X))'$,

$$\|\theta g\| = \lim_{n\to\infty}\|\theta g_n\| = \lim_{n\to\infty}\|g_n\| = \|g\|.$$

This proves the theorem in case $p=1$ and shows θ is the desired isometry.

Next suppose $p=1$ and $g\in L^\infty(\Omega;X')$. It is still the case that $\|\theta g\| \leq \|g\|_{L^\infty(\Omega;X')}$. As above, one must choose f appropriately. In this case, assume μ is a finite measure. Begin with g a simple function $g(\omega)=\sum_{i=1}^m c_i^*\mathcal{X}_{E_i}(\omega)$. Suppose $\|c_1^*\|$ is at least as large as all other $\|c_i^*\|$ modify if the largest of these occurs at $k\neq 1$. Thus $\|g\|_\infty = \|c_i^*\|_{X'}$. Now let $c_1^*(d_1)\geq \|c_1^*\|-\varepsilon\mu(E_i), \|d_1\|_X=1$, and let $f(\omega)\equiv \frac{d_1}{\mu(E_i)}\mathcal{X}_{E_i}(\omega)$. Then $\int_\Omega \|f\|\,d\mu = 1$. Also

$$g(\omega)(f(\omega))\frac{c_1^*(d_1)}{\mu(E_i)}\mathcal{X}_{E_i}(\omega) \geq \frac{1}{\mu(E_i)}\mathcal{X}_{E_i}(\omega)(\|c_1^*\|-\varepsilon\mu(E_i))$$

and so

$$\begin{aligned}
|\theta g(f)| &= \left|\int_\Omega g(\omega)(f(\omega))\,d\mu\right| \geq \left|\int_\Omega \left(\frac{1}{\mu(E_i)}\mathcal{X}_{E_i}(\omega)(\|c_1^*\|-\varepsilon\mu(E_i))\right)d\mu\right|\\
&\geq \|c_1^*\|-\varepsilon\mu(\Omega) = \|g\|_\infty - \varepsilon\mu(\Omega)
\end{aligned}$$

Thus

$$\|g\|_\infty \geq \|\theta g\| \geq \|g\|_\infty - \varepsilon\mu(\Omega)$$

and so, since ε is arbitrary, it follows that $\|\theta g\| = \|g\|_{L^\infty(\Omega;X')}$. Extending from simple functions to functions in $L^\infty(\Omega;X')$ goes as before. Approximate with simple functions and pass to a limit. ■

Theorem 24.8.3 *If X is a Banach space and X' has the Radon Nikodym property, then if $(\Omega,\mathscr{S},\mu)$ is a finite measure space,*

$$(L^p(\Omega;X))' \cong L^{p'}(\Omega;X')$$

and in fact the mapping θ of Theorem 24.8.2 is onto.

Proof: Let $l \in (L^p(\Omega;X))'$ and define $F(E) \in X'$ by

$$F(E)(x) \equiv l(\mathscr{X}_E(\cdot)x).$$

Lemma 24.8.4 *F defined above is a vector measure with values in X' and $|F|(\Omega) < \infty$.*

Proof of the lemma: Clearly $F(E)$ is linear. Also

$$\|F(E)\| = \sup_{\|x\|\le 1} \|F(E)(x)\|$$

$$\le \|l\| \sup_{\|x\|\le 1} \|\mathscr{X}_E(\cdot)x\|_{L^p(\Omega;X)} \le \|l\|\mu(E)^{1/p}.$$

Let $\{E_i\}_{i=1}^{\infty}$ be a sequence of disjoint elements of $\mathscr{S}$ and let $E = \cup_{n<\infty} E_n$.

$$\begin{aligned} \left|F(E)(x) - \sum_{k=1}^{n} F(E_k)(x)\right| &= \left|l(\mathscr{X}_E(\cdot)x) - \sum_{i=1}^{n} l(\mathscr{X}_{E_i}(\cdot)x)\right| \\ &\le \|l\| \left\|\mathscr{X}_E(\cdot)x - \sum_{i=1}^{n} \mathscr{X}_{E_i}(\cdot)x\right\|_{L^p(\Omega;X)} \\ &\le \|l\|\mu\left(\bigcup_{k>n} E_k\right)^{1/p} \|x\|. \end{aligned} \tag{24.36}$$

Since $\mu(\Omega) < \infty$, $\lim_{n\to\infty} \mu\left(\bigcup_{k>n} E_k\right)^{1/p} = 0$ and so inequality 24.36 shows that

$$\lim_{n\to\infty} \left\|F(E) - \sum_{k=1}^{n} F(E_k)\right\|_{X'} = 0.$$

To show $|F|(\Omega) < \infty$, let $\varepsilon > 0$ be given, let $\{H_1,\cdots,H_n\}$ be a partition of Ω, and let $\|x_i\| \le 1$ be chosen in such a way that

$$F(H_i)(x_i) > \|F(H_i)\| - \varepsilon/n.$$

Thus

$$-\varepsilon + \sum_{i=1}^{n} \|F(H_i)\| < \sum_{i=1}^{n} |l(\mathscr{X}_{H_i}(\cdot)x_i)| \le \|l\| \left\|\sum_{i=1}^{n} \mathscr{X}_{H_i}(\cdot)x_i\right\|_{L^p(\Omega;X)}$$

$$\le \|l\| \left(\int_\Omega \sum_{i=1}^{n} \mathscr{X}_{H_i}(s)\,d\mu\right)^{1/p} = \|l\|\mu(\Omega)^{1/p}.$$

Since $\varepsilon > 0$ was arbitrary, $\sum_{i=1}^{n} \|F(H_i)\| < \|l\| \mu(\Omega)^{1/p}$.Since the partition was arbitrary, this shows $|F|(\Omega) \leq \|l\| \mu(\Omega)^{1/p}$ and this proves the lemma. ■

Continuing with the proof of Theorem 24.8.3, note that $F \ll \mu$. Since X' has the Radon Nikodym property, there exists $g \in L^1(\Omega; X')$ such that

$$F(E) = \int_E g(s)\, d\mu.$$

Also, from the definition of $F(E)$,

$$l\left(\sum_{i=1}^{n} x_i \mathcal{X}_{E_i}(\cdot)\right) = \sum_{i=1}^{n} l\left(\mathcal{X}_{E_i}(\cdot) x_i\right)$$

$$= \sum_{i=1}^{n} F(E_i)(x_i) = \sum_{i=1}^{n} \int_{E_i} g(s)(x_i)\, d\mu. \tag{24.37}$$

It follows from 24.37 that whenever h is a simple function,

$$l(h) = \int_\Omega g(s)(h(s))\, d\mu. \tag{24.38}$$

Let $G_n \equiv \{s : \|g(s)\|_{X'} \leq n\}$ and let $j : L^p(G_n; X) \to L^p(\Omega; X)$ be given by

$$jh(s) = \begin{cases} h(s) \text{ if } s \in G_n, \\ 0 \text{ if } s \notin G_n. \end{cases}$$

Letting h be a simple function in $L^p(G_n; X)$,

$$j^* l(h) = l(jh) = \int_{G_n} g(s)(h(s))\, d\mu. \tag{24.39}$$

Since the simple functions are dense in $L^p(G_n; X)$, and $g \in L^{p'}(G_n; X')$, it follows 24.39 holds for all $h \in L^p(G_n; X)$. By Theorem 24.8.2,

$$\|g\|_{L^{p'}(G_n; X')} = \|j^* l\|_{(L^p(G_n; X))'} \leq \|l\|_{(L^p(\Omega; X))'}.$$

By the monotone convergence theorem,

$$\|g\|_{L^{p'}(\Omega; X')} = \lim_{n \to \infty} \|g\|_{L^{p'}(G_n; X')} \leq \|l\|_{(L^p(\Omega; X))'}.$$

Therefore $g \in L^{p'}(\Omega; X')$ and since simple functions are dense in $L^p(\Omega; X)$, 24.38 holds for all $h \in L^p(\Omega; X)$. Thus $l = \theta g$ and the theorem is proved because, by Theorem 24.8.2, $\|l\| = \|g\|$ and so the mapping θ is onto because l was arbitrary. ■

As in the scalar case, everything generalizes to the case of σ finite measure spaces. The proof is almost identical.

Lemma 24.8.5 *Let $(\Omega, \mathcal{S}, \mu)$ be a σ finite measure space and let X be a Banach space such that X' has the Radon Nikodym property. Then there exists a measurable function, r such that $r(x) > 0$ for all x, such that $|r(x)| < M$ for all x, and $\int r d\mu < \infty$. For*

$$\Lambda \in (L^p(\Omega; X))',\ p \geq 1,$$

there exists a unique $h \in L^{p'}(\Omega;X')$, $L^{\infty}(\Omega;X')$ *if* $p=1$ *such that*

$$\Lambda f = \int h(f)\, d\mu.$$

Also $\|h\| = \|\Lambda\|$. ($\|h\| = \|h\|_{p'}$ *if* $p>1$, $\|h\|_{\infty}$ *if* $p=1$). *Here*

$$\frac{1}{p} + \frac{1}{p'} = 1.$$

Proof: First suppose r exists as described. Also, to save on notation and to emphasize the similarity with the scalar case, denote the norm in the various spaces by $|\cdot|$. Define a new measure $\widetilde{\mu}$, according to the rule

$$\widetilde{\mu}(E) \equiv \int_E r d\mu. \tag{24.40}$$

Thus $\widetilde{\mu}$ is a finite measure on $\mathscr{S}$. Now define a mapping, $\eta : L^p(\Omega;X,\mu) \to L^p(\Omega;X,\widetilde{\mu})$ by $\eta f = r^{-\frac{1}{p}} f$. Then

$$\|\eta f\|^p_{L^p(\widetilde{\mu})} = \int \left| r^{-\frac{1}{p}} f \right|^p r d\mu = \|f\|^p_{L^p(\mu)}$$

and so η is one to one and in fact preserves norms. I claim that also η is onto. To see this, let $g \in L^p(\Omega;X,\widetilde{\mu})$ and consider the function, $r^{\frac{1}{p}} g$. Then

$$\int \left| r^{\frac{1}{p}} g \right|^p d\mu = \int |g|^p r d\mu = \int |g|^p d\widetilde{\mu} < \infty$$

Thus $r^{\frac{1}{p}} g \in L^p(\Omega;X,\mu)$ and $\eta\left(r^{\frac{1}{p}} g\right) = g$ showing that η is onto as claimed. Thus η is one to one, onto, and preserves norms. Consider the diagram below which is descriptive of the situation in which η^* must be one to one and onto.

$$\begin{array}{ccccc} & & & \eta^* & \\ h, L^{p'}(\widetilde{\mu}) & & L^p(\widetilde{\mu})', \widetilde{\Lambda} & \to & L^p(\mu)', \Lambda \\ & & & \eta & \\ & & L^p(\widetilde{\mu}) & \leftarrow & L^p(\mu) \end{array}$$

Then for $\Lambda \in L^p(\mu)'$, there exists a unique $\widetilde{\Lambda} \in L^p(\widetilde{\mu})'$ such that $\eta^* \widetilde{\Lambda} = \Lambda, \left\|\widetilde{\Lambda}\right\| = \|\Lambda\|$. By the Riesz representation theorem for finite measure spaces, there exists a unique $h \in L^{p'}(\widetilde{\mu}) \equiv L^{p'}(\Omega;X',\widetilde{\mu})$ which represents $\widetilde{\Lambda}$ in the manner described in the Riesz representation theorem. Thus $\|h\|_{L^{p'}(\widetilde{\mu})} = \left\|\widetilde{\Lambda}\right\| = \|\Lambda\|$ and for all $f \in L^p(\mu)$,

$$\begin{aligned} \Lambda(f) &= \eta^* \widetilde{\Lambda}(f) \equiv \widetilde{\Lambda}(\eta f) = \int h(\eta f)\, d\widetilde{\mu} = \int r h \left(r^{-\frac{1}{p}} f \right) d\mu \\ &= \int r^{\frac{1}{p'}} h f d\mu. \end{aligned}$$

Now

$$\int \left| r^{\frac{1}{p'}} h \right|^{p'} d\mu = \int |h|^{p'} r d\mu = \|h\|^{p'}_{L^{p'}(\widetilde{\mu})} < \infty.$$

Thus $\left\| r^{\frac{1}{p'}} h \right\|_{L^{p'}(\mu)} = \|h\|_{L^{p'}(\widetilde{\mu})} = \left\| \widetilde{\Lambda} \right\| = \|\Lambda\|$ and represents Λ in the appropriate way. If $p = 1$, then $1/p' \equiv 0$. Now consider the existence of r. Since the measure space is σ finite, there exist $\{\Omega_n\}$ disjoint, each having positive measure and their union equals Ω. Then define

$$r(\omega) \equiv \sum_{n=1}^{\infty} \frac{1}{n^2} \mu(\Omega_n)^{-1} \mathcal{X}_{\Omega_n}(\omega)$$

This proves the Lemma.

Theorem 24.8.6 *(Riesz representation theorem) Let $(\Omega, \mathcal{S}, \mu)$ be σ finite and let X' have the Radon Nikodym property. Then for*

$$\Lambda \in (L^p(\Omega; X, \mu))', \ p \geq 1$$

there exists a unique $h \in L^q(\Omega, X', \mu)$, $L^\infty(\Omega, X', \mu)$ if $p = 1$ such that

$$\Lambda f = \int h(f)\, d\mu.$$

Also $\|h\| = \|\Lambda\|$. ($\|h\| = \|h\|_q$ if $p > 1$, $\|h\|_\infty$ if $p = 1$). Here

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Proof: The above lemma gives the existence part of the conclusion of the theorem. Uniqueness is done as before.

Corollary 24.8.7 *If X' is separable, then for $(\Omega, \mathcal{S}, \mu)$ a σ finite measure space,*

$$(L^p(\Omega; X))' \cong L^{p'}(\Omega; X').$$

Corollary 24.8.8 *If X is separable and reflexive, then for $(\Omega, \mathcal{S}, \mu)$ a σ finite measure space,*

$$(L^p(\Omega; X))' \cong L^{p'}(\Omega; X').$$

Corollary 24.8.9 *If X is separable and reflexive and $(\Omega, \mathcal{S}, \mu)$ a σ finite measure space,then if $p \in (1, \infty)$, then $L^p(\Omega; X)$ is reflexive.*

Proof: This is just like the scalar valued case.

24.9 An Example of Polish Space

Here is an interesting example. Obviously $L^\infty(0, T, H)$ is not separable with the normed topology. However, bounded sets turn out to be metric spaces which are complete and separable. This is the next lemma. Recall that a Polish space is a complete separable metric space. In this example, H is a separable real Hilbert space or more generally a separable real Banach space.

Lemma 24.9.1 *Let $B = \overline{B(\mathbf{0}, L)}$ be a closed ball in $L^\infty(0, T, H)$. Then B is a Polish space with respect to the weak $*$ topology. The closure is taken with respect to the usual topology.*

Proof: Let $\{z_k\}_{k=1}^{\infty} = X$ be a dense countable subspace in $L^1(0,T,H)$. You start with a dense countable set and then consider all finite linear combinations having coefficients in $\mathbb{Q}$. Then the metric on B is

$$d(f,g) \equiv \sum_{k=1}^{\infty} 2^{-k} \frac{\left|\langle f-g, z_k\rangle_{L^\infty,L^1}\right|}{1+\left|\langle f-g, z_k\rangle_{L^\infty,L^1}\right|}$$

Is B complete? Suppose you have a Cauchy sequence $\{f_n\}$. This happens if and only if $\{\langle f_n, z_k\rangle\}_{n=1}^{\infty}$ is a Cauchy sequence for each k. Therefore, there exists

$$\xi(z_k) = \lim_{n\to\infty} \langle f_n, z_k\rangle.$$

Then for $a,b \in \mathbb{Q}$, and $z, w \in X$

$$\xi(az+bw) = \lim_{n\to\infty} \langle f_n, az+bw\rangle = \lim_{n\to\infty} a\langle f_n, z\rangle + b\langle f_n, w\rangle = a\xi(z) + b\xi(w)$$

showing that ξ is linear on X a dense subspace of $L^1(0,T,H)$. Is ξ bounded on this dense subspace with bound L? For $z \in X$,

$$|\xi(z)| \equiv \lim_{n\to\infty} |\langle f_n, z\rangle| \le \limsup_{n\to\infty} \|f_n\|_{L^\infty} \|z\|_{L^1} \le L\|z\|_{L^1}$$

Hence ξ is also bounded on this dense subset of $L^1(0,T,H)$. Therefore, there is a unique bounded linear extension of ξ to all of $L^1(0,T,H)$ still denoted as ξ such that its norm in $L^1(0,T,H)'$ is no larger than L. It follows from the Riesz representation theorem that there exists a unique $f \in L^\infty(0,T,H)$ such that for all $w \in L^1(0,T,H)$, $\xi(w) = \langle f, w\rangle$ and $\|f\| \le L$. This f is the limit of the Cauchy sequence $\{f_n\}$ in B. Thus B is complete.

Is B separable? Let $f \in B$. Let $\varepsilon > 0$ be given. Choose M such that $\sum_{k=M+1}^{\infty} 2^{-k} < \frac{\varepsilon}{4}$. Then the finite set $\{z_1, \cdots, z_M\}$ is uniformly integrable. There exists $\delta > 0$ such that if $m(S) < \delta$, then $\int_S |z_k|_H \, dm < \left(\frac{\varepsilon}{4(1+\|f\|_{L^\infty})}\right)$ Then there is a sequence of simple functions $\{s_n\}$ which converge uniformly to f off a set of measure zero, N, $\|s_n\|_{L^\infty} \le \|f\|_{L^\infty}$. By regularity of the measure, there exists a continuous function with compact support h_n such that $s_n = h_n$ off a set of measure no more than $\delta/4^n$ and also $\|h_n\|_{L^\infty} \le \|f\|_{L^\infty}$. Then off a set of measure no more than $\frac{1}{3}\delta$, $h_n(r) \to f(r)$. Now by Eggorov's theorem and outer regularity, one can enlarge this exceptional set to obtain an open set S of measure no more than $\delta/2$ such that the convergence is uniform off this exceptional set. Thus f equals the uniform limit of continuous functions on S^C. Define

$$h(r) \equiv \begin{cases} \lim_{n\to\infty} h_n(r) = f(r) \text{ on } S^C \\ \mathbf{0} \text{ on } S\setminus N \\ \mathbf{0} \text{ on } N \end{cases}$$

Then $\|h\|_{L^\infty} \le \|f\|_{L^\infty}$. Now consider $\bar{h} * \psi_m(r)$ where ψ_r is approximate identity.

$$\begin{aligned} \psi_m(t) &= \frac{1}{2} m \mathcal{X}_{[-1/m,1/m]}(t), \ \bar{h} * \psi_m(t) \\ &= \frac{1}{2} m \int_{-1/m}^{1/m} \bar{h}(t-s)\, ds = \frac{1}{2} m \int_{t-1/m}^{t+1/m} \bar{h}(s)\, ds \end{aligned}$$

where we define $\bar{h}$ to be the 0 extension of $\bar{h}$ off $[0,T]$. This is a continuous function of t. Also *a.e.t* is a Lebesgue point and so for *a.e.t*,

$$\left|\frac{1}{2}m\int_{t-1/m}^{t+1/m}\bar{\boldsymbol{h}}(s)\,ds-\bar{\boldsymbol{h}}(t)\right|\to 0$$

$$\left|\bar{\boldsymbol{h}}*\psi_m(r)\right|\equiv\left|\int_{\mathbb{R}}\bar{\boldsymbol{h}}(r-s)\,\psi_m(s)\,ds\right|\leq\|\boldsymbol{h}\|_{L^\infty}\leq\|\boldsymbol{f}\|_{L^\infty}$$

Thus this continuous function is in $L^\infty(0,T,H)$. Letting $\boldsymbol{z}=\boldsymbol{z}_k\in L^1(0,T,H)$ be one of those defined above,

$$\left|\int_0^T\left\langle\bar{\boldsymbol{h}}*\psi_m(t)-\boldsymbol{f}(t),\boldsymbol{z}(t)\right\rangle dt\right|\leq\int_0^T\left|\left\langle\bar{\boldsymbol{h}}*\psi_m(t)-\boldsymbol{h}(t),\boldsymbol{z}(t)\right\rangle\right|dt$$

$$+\int_0^T\left|\left\langle\boldsymbol{h}(t)-\boldsymbol{f}(t),\boldsymbol{z}(t)\right\rangle\right|dt \tag{24.41}$$

for a.e. $t,\bar{\boldsymbol{h}}*\psi_m(t)-\boldsymbol{h}(t)\to 0$ and the integrand in the first integral in the above is bounded by $2\|\boldsymbol{f}\|_{L^\infty}|\boldsymbol{z}(t)|_H$ so by the dominated convergence theorem, as $m\to\infty$, the first integral converges to 0. As to the second, it is dominated by

$$\int_S\left|\left\langle\boldsymbol{h}(t)-\boldsymbol{f}(t),\boldsymbol{z}(t)\right\rangle\right|dt\leq 2\|\boldsymbol{f}\|_{L^\infty}\int_S|\boldsymbol{z}(t)|\,dt<\frac{2\|\boldsymbol{f}\|_{L^\infty}\varepsilon}{4(1+\|\boldsymbol{f}\|_{L^\infty})}\leq\frac{\varepsilon}{2}$$

Therefore, choosing m large enough so that the first integral on the right in 24.41 is less than $\frac{\varepsilon}{4}$ for each $\boldsymbol{z}_k$ for $k\leq M$, then for each of these,

$$\begin{aligned}d\left(\boldsymbol{f},\bar{\boldsymbol{h}}*\psi_m\right)&\leq\frac{\varepsilon}{4}+\sum_{k=1}^M 2^{-k}\frac{(\varepsilon/4)+(\varepsilon/2)}{1+((\varepsilon/4)+(\varepsilon/2))}=\frac{\varepsilon}{4}+\sum_{k=1}^M 2^{-k}\frac{3}{4}\frac{\varepsilon}{\frac{3}{4}\varepsilon+1}\\&\leq\frac{\varepsilon}{4}+\frac{3\varepsilon}{4}\sum_{k=1}^M 2^{-k}<\frac{\varepsilon}{4}+\frac{3\varepsilon}{4}=\varepsilon\end{aligned}$$

which appears to show that $C([0,T],H)$ is weak $*$ dense in $L^\infty(0,T,H)$. However, this last space is obviously separable in terms of the norm topology. Let D be a countable dense subset of $C([0,T],H)$. For $\boldsymbol{f}\in L^\infty(0,T,H)$ let $\boldsymbol{g}\in C([0,T],H)$ such that $d(\boldsymbol{f},\boldsymbol{g})<\frac{\varepsilon}{4}$. Then let $\boldsymbol{h}\in D$ be so close to $\boldsymbol{g}$ in $C([0,T],H)$ that

$$\sum_{k=1}^M 2^{-k}\frac{\left|\langle\boldsymbol{h}-\boldsymbol{g},\boldsymbol{z}_k\rangle_{L^\infty,L^1}\right|}{1+\left|\langle\boldsymbol{h}-\boldsymbol{g},\boldsymbol{z}_k\rangle_{L^\infty,L^1}\right|}<\frac{\varepsilon}{2}$$

Then $d(\boldsymbol{f},\boldsymbol{h})\leq d(\boldsymbol{f},\boldsymbol{g})+d(\boldsymbol{g},\boldsymbol{h})<\frac{\varepsilon}{4}+\frac{\varepsilon}{2}+\frac{\varepsilon}{4}=\varepsilon$ It appears that D is dense in B in the weak $*$ topology. ■

24.10 Weakly Convergent Sequences

There is an interesting little result which relates to weak limits in $L^2(\Gamma,E)$ for E a Banach space. I am not sure where to put this thing but think that this would be a good place for it. It obviously generalizes to L^p spaces.

Proposition 24.10.1 *Let E be a Banach space and let $\{u_n\}$ be a sequence in $L^2(\Gamma,E)$ and let $G(x)$ be a weakly compact set in E, and $u_n(x) \in G(x)$ a.e. for each n. Let $\limsup\{u_n(x)\}$ denote the set of all weak limits of subsequences of $\{u_n(x)\}$ and let $H(x)$ be the closure of the convex hull of $\limsup\{u_n(x)\}$. Then if $u_n \to u$ weakly in $L^2(\Gamma,E)$, then $u(x) \in H(x)$ for a.e. x.*

Proof: Let $H = \left\{w \in L^2(\Gamma,E) : w(x) \in H(x) \text{ a.e.}\right\}$. Then H is convex. If you have $w_i \in H$, then since each $H(x)$ is convex, it follows that $\lambda w_1(x) + (1-\lambda) w_2(x) \in H$ for a.e. x and $\lambda \in [0,1]$. Is H closed? Suppose you have $w_n \in H$ and $w_n \to w$ in $L^2(\Gamma,E)$. Then there is a subsequence such that pointwise convergence happens a.e. and so since H is closed, you have $w(x) \in H$ for a.e. x. Hence H is also weakly closed in $L^2(\Gamma,H)$. Thus if u is the weak limit of $\{u_n\}$ in $L^2(\Gamma,E)$, it must be the case that $u(x) \in H(x)$ a.e. ■

As a case of this which might be pretty interesting, suppose $G(x)$ is not just weakly compact but also convex. Then $H(x) = G(x)$ and you can say that $u(x) \in H(x)$ a.e. whenever it is a weak limit in $L^2(\Gamma,E)$ of functions u_n for which $u_n(x) \in G(x)$.

24.11 Some Embedding Theorems

The next lemma is a very useful little result which involves embeddings of Banach spaces.

Lemma 24.11.1 *Suppose $V \subseteq W$ and the injection map is compact, hence continuous. Suppose also that $W \subseteq U$ with continuous injection. Then for any $\varepsilon > 0$ there exists C_ε such that for all $v \in V$,*

$$\|v\|_W \leq \varepsilon \|v\|_V + C_\varepsilon \|v\|_U$$

Proof: Suppose not. Then there exists $\varepsilon > 0$ for which things don't work out. Thus there exists $v_n \in V$ such that

$$\|v_n\|_W > \varepsilon \|v_n\|_V + n \|v_n\|_U$$

Dividing by $\|v_n\|_V$, it can also be assumed that $\|v_n\|_V = 1$. Thus

$$\|v_n\|_W > \varepsilon + n \|v_n\|_U$$

and so $\|v_n\|_U \to 0$. However, v_n is contained in the closed unit ball of V which is, by assumption precompact in W. Hence, there exists a subsequence, still denoted as $\{v_n\}$ such that $v_n \to v$ in W. But it was just determined that $v = 0$ and so

$$0 \geq \limsup_{n\to\infty} (\varepsilon + n \|v_n\|_U) \geq \varepsilon$$

which is a contradiction. ■

Recall the following definition, this time for the space of continuous functions defined on a compact set with values in a Banach space.

Definition 24.11.2 *Let $\mathscr{A} \subseteq C(K;V)$ where the last symbol denotes the continuous functions defined on a compact set $K \subseteq X$ a metric space having values in V a Banach space. Then $\mathscr{A}$ is equicontinuous if for every $\varepsilon > 0$, there exists $\delta > 0$ such that for every $f \in \mathscr{A}$, if $d(x,y) < \delta$, then*

$$\|f(x) - f(y)\|_V < \varepsilon.$$

Also $\mathscr{A} \subseteq C(K;V)$ is uniformly bounded means

$$\sup_{f\in\mathscr{A}} \|f\|_{\infty,V} < \infty \text{ where } \|f\|_{\infty,V} \equiv \max_{x\in K} \|f(x)\|_V.$$

Here is a general version of the Ascoli Arzela theorem valid for Banach spaces.

Theorem 24.11.3 *Let $V \subseteq W \subseteq U$ where the injection map of V into W is compact and W embedds continuously into U, these being Banach spaces. Assume:*

1. *$\mathscr{A} \subseteq C(K;U)$ where K is compact and $\mathscr{A}$ is equicontinuous.*
2. *$\sup_{f\in\mathscr{A}} \|f\|_{\infty,V} < \infty$ where $\|f\|_{\infty,V} \equiv \max_{x\in K} \|f(x)\|_V$.*

Then

1. *$\mathscr{A} \subseteq C(K;W)$ and $\mathscr{A}$ is equicontinuous into W*
2. *$\mathscr{A}$ is pre-compact in $C(K;W)$. This means that $\overline{\mathscr{A}}$ is compact in $C(K;W)$.*

Proof: Let $C \equiv \sup_{f\in\mathscr{A}} \|f\|_{\infty,V} < \infty$. Let $\varepsilon > 0$ be given. Then from Lemma 24.11.1,

$$\|f(x) - f(y)\|_W \leq \frac{\varepsilon}{5C} \|f(x) - f(y)\|_V + C_\varepsilon \|f(x) - f(y)\|_U$$

$$\leq \frac{2\varepsilon}{5} + C_\varepsilon \|f(x) - f(y)\|_U$$

By equicontinuity in $C(K,U)$, there exists a $\delta > 0$ such that if $d(x,y) < \delta$, then for all $f \in \mathscr{A}$,

$$C_\varepsilon \|f(x) - f(y)\|_U < \frac{2\varepsilon}{5}$$

Thus if $d(x,y) < \delta$, then $\|f(x) - f(y)\|_W < \varepsilon$ for all $f \in \mathscr{A}$.

It remains to verify that $\mathscr{A}$ is pre-compact in $C(K;W)$. Since this space of continuous functions is complete, it suffices to verify that for all $\varepsilon > 0$, $\mathscr{A}$ has an ε net. Suppose then that for some $\varepsilon > 0$ there is no ε net. Thus there is an infinite sequence $\{f_n\}$ for which $\|f_n - f_m\|_{\infty,W} \geq \varepsilon$ whenever $m \neq n$. There exists $\delta > 0$ such that if $d(x,y) < \delta$, then for all f_n,

$$\|f_n(x) - f_n(y)\|_W < \frac{\varepsilon}{5}.$$

Let $\{x_k\}_{k=1}^p$ be a $\delta/2$ net for K. This is where we use K is compact. By compactness of the embedding of V into W, there exists a further subsequence, still called $\{f_n\}$ such that each $\{f_n(x_k)\}_{n=1}^\infty$ converges, this for each x_k in that $\delta/2$ net. Thus there is a single N such that if $n > N$, then for all $m,n > N$, and $k \leq p$,

$$\|f_n(x_k) - f_m(x_k)\|_W < \frac{\varepsilon}{5}$$

Now letting $x \in K$ be arbitrary, it is in $B(x_k, \delta/2)$ for some x_k. Therefore, for n,m larger than N,

$$\begin{aligned} \|f_n(x) - f_m(x)\|_W &\leq \|f_n(x) - f_n(x_k)\|_W \\ &\quad + \|f_n(x_k) - f_m(x_k)\|_W \\ &\quad + \|f_m(x_k) - f_m(x)\| \end{aligned}$$

$$< \frac{\varepsilon}{5} + \frac{\varepsilon}{5} + \frac{\varepsilon}{5} = \frac{3\varepsilon}{5}$$

Taking the maximum for all x, for $m,n > N$,

$$\|f_n - f_m\|_{W,\infty} \leq \frac{3\varepsilon}{5} < \varepsilon$$

contrary to the assumption that every pair is further apart than ε. Thus $\mathscr{A}$ is totally bounded so its closure would also be totally bounded and complete. In other words, $\mathscr{A}$ is precompact in $C(K;W)$. ■

In the following theorem about compact subsets of an L^p space, the measure will be Lebesgue measure. It depends on the above version of the Ascoli Arzela theorem. First note the following which I will use when convenient. For $a,b \geq 0$, and $p \geq 1$,

$$(a+b)^p \leq 2^{p-1}(a^p + b^p)$$

This follows from the convexity of $y = x^p$ for $x > 0$. Also, for such p,

$$(a+b)^{1/p} \leq a^{1/p} + b^{1/p}$$

Usually the thing of interest in this theorem is the case where $V = W = U = \mathbb{R}$. However, the more general version to be presented is interesting I think. Of course closed and bounded sets are compact in $\mathbb{R}$ so the usual case works as a special case of what is about to be presented.

Theorem 24.11.4 *Let $V \subseteq W \subseteq U$ where these are Banach spaces such that the injection map of V into W is compact and the injection map of W into U is continuous. Let Ω be an open set in $\mathbb{R}^m$ and let $\mathscr{A}$ be a bounded subset of $L^p(\Omega;V)$ and suppose that for all $\varepsilon > 0$, there exist a $\delta > 0$ such that if $|\boldsymbol{h}| < \delta$, then for $\tilde{u}$ denoting the zero extension of u off Ω,*

$$\int_{\mathbb{R}^m} \|\widetilde{u}(\boldsymbol{x}+\boldsymbol{h}) - \widetilde{u}(\boldsymbol{x})\|_U^p \, dx < \varepsilon^p \tag{24.42}$$

Suppose also that for each $\varepsilon > 0$ there exists an open set, $G_\varepsilon \subseteq \Omega$ such that $\overline{G_\varepsilon} \subseteq \Omega$ is compact and for all $u \in \mathscr{A}$,

$$\int_{\Omega \setminus \overline{G_\varepsilon}} \|u(\boldsymbol{x})\|_W^p \, dx < \varepsilon^p \tag{24.43}$$

Then $\mathscr{A}$ is precompact in $L^p(\mathbb{R}^n;W)$.

Proof: Let $\infty > M \geq \sup_{u \in L^p(\Omega;V)} \|u\|_{L^p(\Omega;V)}^p$. Let $\{\psi_n\}$ be a mollifier with support in $B(\mathbf{0}, 1/n)$. I need to show that $\mathscr{A}$ has an η net in $L^p(\Omega;W)$ for every $\eta > 0$. Suppose for some $\eta > 0$ it fails to have an η net. Without loss of generality, let $\eta < 1$. Then by 24.43, it follows that for small enough $\varepsilon > 0$, $\mathscr{A}_\varepsilon \equiv \left\{ u\mathcal{X}_{\overline{G_\varepsilon}} : u \in \mathscr{A} \right\}$ fails to have an $\eta/2$ net. Indeed, pick ε small enough that for all $u \in \mathscr{A}$,

$$\left\| u\mathcal{X}_{\overline{G_\varepsilon}} - u \right\|_{L^p(\Omega;W)} < \frac{\eta}{5}$$

Then if $\left\{ u_k \mathcal{X}_{\overline{G_\varepsilon}} \right\}_{k=1}^r$ is an $\eta/2$ net for $\mathscr{A}_\varepsilon$, so that $\cup_{k=1}^r B\left(u_k \mathcal{X}_{\overline{G_\varepsilon}}, \frac{\eta}{2} \right) \supseteq \mathscr{A}_\varepsilon$, then for $w \in \mathscr{A}$, $w\mathcal{X}_{\overline{G_\varepsilon}} \in B\left(u_k \mathcal{X}_{\overline{G_\varepsilon}}, \frac{\eta}{2} \right)$ for some u_k. Hence,

$$\begin{aligned} \|w - u_k\|_{L^p(\Omega;W)} &\leq \left\| w - w\mathcal{X}_{\overline{G_\varepsilon}} \right\|_{L^p(\Omega;W)} + \left\| w\mathcal{X}_{\overline{G_\varepsilon}} - u_k \mathcal{X}_{\overline{G_\varepsilon}} \right\|_{L^p(\Omega;W)} \\ &\quad + \left\| u_k \mathcal{X}_{\overline{G_\varepsilon}} - u_k \right\|_{L^p(\Omega;W)} \end{aligned}$$

$$\leq \frac{\eta}{5}+\frac{\eta}{2}+\frac{\eta}{5}<\eta$$

and so $\{u_k\}_{k=1}^r$ would be an η net for $\mathscr{A}$ which is assumed to not exist.

Pick this ε in all that follows. By compactness, Lemma 24.11.1, there exists C_η such that for all $u \in V$,

$$\|u\|_W^p \leq \frac{\eta}{50\,(2^{p-1})\,M}\|u\|_V^p + C_\eta\|u\|_U^p \tag{24.44}$$

Let $\mathscr{A}_{\varepsilon n}$ consist of $\mathscr{A}_{\varepsilon n} \equiv \left\{u\mathscr{X}_{\overline{G_\varepsilon}} * \psi_n : u \in \mathscr{A}\right\}$. I want to show that $\mathscr{A}_{\varepsilon n}$ satisfies the conditions for Theorem 24.11.3.

Lemma 24.11.5 *For each n, $\mathscr{A}_{\varepsilon n}$ satisfies the conditions of Theorem 24.11.3.*

Proof: First consider the equicontinuity condition of that theorem. It suffices to show that if $\eta > 0$ then there exists $\delta > 0$ such that if $|\boldsymbol{h}| < \delta$, then for any $u \in \mathscr{A}$ and $\boldsymbol{x} \in \overline{G_\varepsilon}$,

$$\left\|u\mathscr{X}_{\overline{G_\varepsilon}} * \psi_n(\boldsymbol{x}+\boldsymbol{h}) - u\mathscr{X}_{\overline{G_\varepsilon}} * \psi_n(\boldsymbol{x})\right\|_U < \eta$$

Always assume $|\boldsymbol{h}| < \operatorname{dist}\left(\overline{G_\varepsilon}, \Omega^C\right)$, and $\boldsymbol{x} \in \overline{G_\varepsilon}$. Also assume that $|\boldsymbol{h}|$ is small enough that

$$\left(\int_{\mathbb{R}^m}\left|\left(\mathscr{X}_{\overline{G_\varepsilon}}(\boldsymbol{x}-\boldsymbol{y}+\boldsymbol{h}) - \mathscr{X}_{\overline{G_\varepsilon}}(\boldsymbol{x}-\boldsymbol{y})\right)\psi_n(\boldsymbol{y})\right|^{p'} dz\right)^{1/p'} =$$

$$\left(\int_{\mathbb{R}^m}\left|\left(\mathscr{X}_{\overline{G_\varepsilon}}(\boldsymbol{z}+\boldsymbol{h}) - \mathscr{X}_{\overline{G_\varepsilon}}(\boldsymbol{z})\right)\psi_n(\boldsymbol{x}-\boldsymbol{z})\right|^{p'} dz\right)^{1/p'} < \frac{\eta}{2M} \tag{24.45}$$

This can be obtained because by Holder's inequality,

$$\left(\int_{\mathbb{R}^m}\left|\left(\mathscr{X}_{\overline{G_\varepsilon}}(\boldsymbol{z}+\boldsymbol{h}) - \mathscr{X}_{\overline{G_\varepsilon}}(\boldsymbol{z})\right)\psi_n(\boldsymbol{x}-\boldsymbol{z})\right|^{p'} dz\right)^{1/p'}$$

$$\leq \left(\int_{\mathbb{R}^m}\left|\mathscr{X}_{\overline{G_\varepsilon}}(\boldsymbol{z}+\boldsymbol{h}) - \mathscr{X}_{\overline{G_\varepsilon}}(\boldsymbol{z})\right|^{2p'} dz\right)^{\frac{1}{2p'}}\left(\int_{\mathbb{R}^m}\psi_n(\boldsymbol{x}-\boldsymbol{z})^{2p'} dz\right)^{\frac{1}{2p'}}$$

which is small independent of $\boldsymbol{x}$ for $|\boldsymbol{h}|$ small enough, thanks to continuity of translation in $L^{2p'}(\mathbb{R}^m)$. Then

$$\begin{aligned}
&\left\|u\mathscr{X}_{\overline{G_\varepsilon}} * \psi_n(\boldsymbol{x}+\boldsymbol{h}) - u\mathscr{X}_{\overline{G_\varepsilon}} * \psi_n(\boldsymbol{x})\right\|_U \\
= &\left\|\int_{\mathbb{R}^m}\left(\tilde{u}(\boldsymbol{x}+\boldsymbol{h}-\boldsymbol{y})\,\mathscr{X}_{\overline{G_\varepsilon}}(\boldsymbol{x}+\boldsymbol{h}-\boldsymbol{y}) - \tilde{u}(\boldsymbol{x}-\boldsymbol{y})\,\mathscr{X}_{\overline{G_\varepsilon}}(\boldsymbol{x}-\boldsymbol{y})\right)\psi_n(\boldsymbol{y})\,dy\right\|_U \\
\leq &\int_{\mathbb{R}^m}\left\|\left(\tilde{u}(\boldsymbol{x}+\boldsymbol{h}-\boldsymbol{y})\,\mathscr{X}_{\overline{G_\varepsilon}}(\boldsymbol{x}+\boldsymbol{h}-\boldsymbol{y}) - \tilde{u}(\boldsymbol{x}-\boldsymbol{y})\,\mathscr{X}_{\overline{G_\varepsilon}}(\boldsymbol{x}-\boldsymbol{y})\right)\right\|_U \psi_n(\boldsymbol{y})\,dy
\end{aligned}$$

Changing the variables,

$$\leq \int_{\mathbb{R}^m}\left\|\begin{array}{c}(\tilde{u}(\boldsymbol{z}+\boldsymbol{h})-\tilde{u}(\boldsymbol{z}))\,\mathscr{X}_{\overline{G_\varepsilon}}(\boldsymbol{z}+\boldsymbol{h}) \\ +\tilde{u}(\boldsymbol{z})\left(\mathscr{X}_{\overline{G_\varepsilon}}(\boldsymbol{z}+\boldsymbol{h}) - \mathscr{X}_{\overline{G_\varepsilon}}(\boldsymbol{z})\right)\end{array}\right\|_U \psi_n(\boldsymbol{x}-\boldsymbol{z})\,dz$$

$$\leq \int_{\mathbb{R}^m} \left\| \left(\tilde{u}(z+h)-\tilde{u}(z)\right) \mathcal{X}_{\overline{G_\varepsilon}}(z+h) \right\|_U \psi_n(x-z)\,dz$$
$$+ \int_{\mathbb{R}^m} \|\tilde{u}(z)\|_U \left| \mathcal{X}_{\overline{G_\varepsilon}}(z+h) - \mathcal{X}_{\overline{G_\varepsilon}}(z) \right| \psi_n(x-z)\,dz \tag{24.46}$$

The first integral

$$\leq \left(\int_{\mathbb{R}^m} \|\tilde{u}(z+h)-\tilde{u}(z)\|_U^p \right)^{1/p} \left(\int_{\mathbb{R}^m} \psi_n^{p'}(x-z)\,dz \right)^{1/p'}$$

You make the obvious change here in case $p=1$. Instead of the above, you would have

$$\leq \int_{\mathbb{R}^m} \|\tilde{u}(z+h)-\tilde{u}(z)\|_U\,dz 2\,\|\psi_n\|_\infty$$

Since Lebesgue measure is translation independent, there is a constant C_n such that the above is

$$\leq C_n \left(\int_{\mathbb{R}^m} \|\tilde{u}(z+h)-\tilde{u}(z)\|_U^p \right)^{1/p} < \eta/2$$

and this holds for all $u \in \mathscr{A}$. As for the second integral in 24.46, from 24.45, it follows that this term is no larger than

$$\leq \left(\int_{\mathbb{R}^m} \|\tilde{u}(z)\|_U^p\,dz \right)^{1/p} \left(\int_{\mathbb{R}^m} \left(\left| \mathcal{X}_{\overline{G_\varepsilon}}(z+h) - \mathcal{X}_{\overline{G_\varepsilon}}(z) \right| \psi_n(x-z) \right)^{p'} dz \right)^{1/p'}$$

and by 24.45,

$$< M\frac{\eta}{2M} = \frac{\eta}{2}$$

Thus, if $\delta < \operatorname{dist}\left(\overline{G_\varepsilon}, \Omega^C\right)$ and 24.45 holds, then for all $u \in \mathscr{A}$, when $|h| < \delta$,

$$\left\| u\mathcal{X}_{\overline{G_\varepsilon}} * \psi_n(x+h) - u\mathcal{X}_{\overline{G_\varepsilon}} * \psi_n(x) \right\|_U < \eta$$

and so the desired equicontinuity condition holds for $\mathscr{A}_{\varepsilon n}$. Note that δ does depend on n but for each n, things work out well.

I also need to verify that the functions in $\mathscr{A}_{\varepsilon n}$ are uniformly bounded. For $x \in \overline{G_\varepsilon}$ and $u \in \mathscr{A}$,

$$\left\| u\mathcal{X}_{\overline{G_\varepsilon}} * \psi_n(x) \right\|_V \leq \int_{\overline{G_\varepsilon}} \|u(z)\|\,\psi_n(x-z)\,dz$$
$$\leq \left(\int_\Omega \|u(z)\|^p\,dz \right)^{1/p} \left(\int_\Omega \psi_n(x-z)^{p'} \right)^{1/p'} \leq MC_n \blacksquare$$

Now is a general statement about norms, indicating that the L^p norm is no more than a constant times the norm involving the maximum.

$$\left(\int_{\overline{G_\varepsilon}} \|v(x)\|_W^p\,dx \right)^{1/p} \leq \max_{x \in \overline{G_\varepsilon}} \|v(x)\|_W\, m\left(\overline{G_\varepsilon}\right) \equiv m\left(\overline{G_\varepsilon}\right) \|v\|_{W,\infty}$$

It follows from Theorem 24.11.3 that for every $\eta > 0$, there exists a η net in $C\left(\overline{G_\varepsilon}; W\right)$ for $\mathscr{A}_{\varepsilon n}$, this for each n. Then from the above inequality, it follows that for each η, there exists an η net in $L^p\left(\overline{G_\varepsilon}; W\right)$ for $\mathscr{A}_{\varepsilon n}$.

Recall also, from the assumption that the theorem is not true,

$$\mathscr{A}_\varepsilon \equiv \left\{ u\mathscr{X}_{\overline{G_\varepsilon}} : u \in \mathscr{A} \right\}$$

has no $\eta/2$ net in $L^p\left(\overline{G_\varepsilon}; W\right)$. Next I estimate the distance in $L^p\left(\overline{G_\varepsilon}; W\right)$ between $u\mathscr{X}_{\overline{G_\varepsilon}}$ for $u \in \mathscr{A}$ and $u\mathscr{X}_{\overline{G_\varepsilon}} * \psi_n$. The idea is that for each $n, \mathscr{A}_{\varepsilon n}$ has an $\eta/8$ net and for n large enough, $u\mathscr{X}_{\overline{G_\varepsilon}}$ is close to $u\mathscr{X}_{\overline{G_\varepsilon}} * \psi_n$ so a contradiction will result if the functions of the second sort are totally bounded while those functions of the first sort don't. Assume always that $1/n < \operatorname{dist}\left(\overline{G_\varepsilon}, \Omega^C\right)$. Using Minkowski's inequality,

$$\left\| u\mathscr{X}_{\overline{G_\varepsilon}} - u\mathscr{X}_{\overline{G_\varepsilon}} * \psi_n \right\|_{L^p\left(\overline{G_\varepsilon}; W\right)} =$$

$$\left(\int_{\mathbb{R}^m} \left\| \int_{\mathbb{R}^m} \left(u\mathscr{X}_{\overline{G_\varepsilon}}(\boldsymbol{x}) - u\mathscr{X}_{\overline{G_\varepsilon}}(\boldsymbol{x}-\boldsymbol{y}) \right) \psi_n(\boldsymbol{y})\, dy \right\|_W^p dx \right)^{1/p}$$

$$\leq \int_{B(\mathbf{0},1/n)} \psi_n(\boldsymbol{y}) \left(\int_{\mathbb{R}^m} \left\| \left(u\mathscr{X}_{\overline{G_\varepsilon}}(\boldsymbol{x}) - u\mathscr{X}_{\overline{G_\varepsilon}}(\boldsymbol{x}-\boldsymbol{y}) \right) \right\|_W^p dx \right)^{1/p} dy$$

$$\leq \int_{B(\mathbf{0},1/n)} \psi_n(\boldsymbol{y}) \left(\int_{\mathbb{R}^m} \left\| \left(\tilde{u}(\boldsymbol{x}) - \tilde{u}(\boldsymbol{x}-\boldsymbol{y}) \right) \right\|_W^p dx \right)^{1/p} dy$$

$$\leq \int_{B(\mathbf{0},1/n)} \psi_n(\boldsymbol{y}) \left(\begin{array}{c} \int_{\mathbb{R}^m} \frac{\eta}{50\left(2^{p-1}\right)M} \left\| \tilde{u}(\boldsymbol{x}) - \tilde{u}(\boldsymbol{x}-\boldsymbol{y}) \right\|_V^p \\ + C_\eta \left\| \tilde{u}(\boldsymbol{x}) - \tilde{u}(\boldsymbol{x}-\boldsymbol{y}) \right\|_U^p dx \end{array} \right)^{1/p} dy$$

$$\leq \int_{B\left(\mathbf{0},\frac{1}{n}\right)} \psi_n(\boldsymbol{y}) \left(\begin{array}{c} \int_{\mathbb{R}^m} \frac{\eta}{50\left(2^{p-1}\right)M} 2^{p-1} 2 \left(\left\| \tilde{u}(\boldsymbol{x}) \right\|_V^p \right) dx \\ + C_\eta \int_{\mathbb{R}^m} \left\| \tilde{u}(\boldsymbol{x}) - \tilde{u}(\boldsymbol{x}-\boldsymbol{y}) \right\|_U^p dx \end{array} \right)^{1/p} dy$$

$$\leq \int_{B\left(\mathbf{0},\frac{1}{n}\right)} \psi_n(\boldsymbol{y}) \left(\begin{array}{c} \int_{\mathbb{R}^m} \frac{\eta}{25M} \left(\left\| \tilde{u}(\boldsymbol{x}) \right\|_V^p \right) dx \\ + \int_{\mathbb{R}^m} C_\eta \left\| \tilde{u}(\boldsymbol{x}) - \tilde{u}(\boldsymbol{x}-\boldsymbol{y}) \right\|_U^p dx \end{array} \right)^{1/p} dy$$

$$\leq \int_{B\left(\mathbf{0},\frac{1}{n}\right)} \psi_n(\boldsymbol{y}) \left(\frac{\eta}{25} + \int_{\mathbb{R}^m} C_\eta \left\| \tilde{u}(\boldsymbol{x}) - \tilde{u}(\boldsymbol{x}-\boldsymbol{y}) \right\|_U^p dx \right)^{1/p} dy$$

By assumption 24.42, there exists N such that if $n \geq N$, then $|\boldsymbol{y}| < \frac{1}{n}$ and for all $u \in \mathscr{A}$,

$$\left\| u\mathscr{X}_{\overline{G_\varepsilon}} - u\mathscr{X}_{\overline{G_\varepsilon}} * \psi_n \right\|_{L^p\left(\overline{G_\varepsilon}; W\right)} \leq$$

$$\begin{aligned} & \int_{B\left(\mathbf{0},\frac{1}{n}\right)} \psi_n(\boldsymbol{y}) \left(\frac{\eta}{25} + \frac{\eta^p}{8^p} \right)^{1/p} dy \\ \leq & \int_{B\left(\mathbf{0},\frac{1}{n}\right)} \psi_n(\boldsymbol{y}) \left(\frac{\eta}{25} + \frac{\eta}{8} \right) dy \\ = & \frac{\eta}{25} + \frac{\eta}{8} \end{aligned}$$

Recall $\eta < 1$.

Let n be this large. Then let $\left\{u_k \mathscr{X}_{\overline{G_\varepsilon}} * \psi_n\right\}_{k=1}^r$ be a $\frac{\eta}{8}$ net for $\mathscr{A}_{\varepsilon n}$ in $L^p\left(\overline{G_\varepsilon}; W\right)$. Then consider the balls $B\left(u_k \mathscr{X}_{\overline{G_\varepsilon}}, \frac{\eta}{4}\right)$ in $L^p\left(\overline{G_\varepsilon}; W\right)$. If $w\mathscr{X}_{\overline{G_\varepsilon}}$ is in $\mathscr{A}_\varepsilon$, is it in some $B\left(u_k \mathscr{X}_{\overline{G_\varepsilon}}, \frac{\eta}{2}\right)$? By what was just shown, there is k such that

$$\left\| w\mathscr{X}_{\overline{G_\varepsilon}} * \psi_n - u_k \mathscr{X}_{\overline{G_\varepsilon}} * \psi_n \right\|_{L^p\left(\overline{G_\varepsilon}; W\right)} < \frac{\eta}{8}$$

and also

$$\left\| w\mathscr{X}_{\overline{G_\varepsilon}} - w\mathscr{X}_{\overline{G_\varepsilon}} * \psi_n \right\|_{L^p\left(\overline{G_\varepsilon}; W\right)} < \frac{\eta}{8} + \frac{\eta}{25}$$

$$\left\| u_k \mathscr{X}_{\overline{G_\varepsilon}} - u_k \mathscr{X}_{\overline{G_\varepsilon}} * \psi_n \right\|_{L^p\left(\overline{G_\varepsilon}; W\right)} < \frac{\eta}{8} + \frac{\eta}{25}$$

Thus,

$$\begin{aligned} \left\| w\mathscr{X}_{\overline{G_\varepsilon}} - u_k \mathscr{X}_{\overline{G_\varepsilon}} \right\|_{L^p(G_\varepsilon; W)} &\leq \left\| w\mathscr{X}_{\overline{G_\varepsilon}} - w\mathscr{X}_{\overline{G_\varepsilon}} * \psi_n \right\|_{L^p\left(\overline{G_\varepsilon}; W\right)} \\ &\quad + \left\| w\mathscr{X}_{\overline{G_\varepsilon}} * \psi_n - u_k \mathscr{X}_{\overline{G_\varepsilon}} * \psi_n \right\|_{L^p\left(\overline{G_\varepsilon}; W\right)} \\ &\quad + \left\| u_k \mathscr{X}_{\overline{G_\varepsilon}} * \psi_n - u_k \mathscr{X}_{\overline{G_\varepsilon}} \right\|_{L^p\left(\overline{G_\varepsilon}; W\right)} \\ &< \frac{3\eta}{8} + \frac{2\eta}{25} < \frac{\eta}{2} \end{aligned}$$

It follows that $\left\{u_k \mathscr{X}_{\overline{G_\varepsilon}}\right\}_{k=1}^r$ is a $\eta/2$ net for $L^p\left(\overline{G_\varepsilon}; W\right)$ contrary to the construction. Thus $\mathscr{A}$ has an η net after all. ■

In case Ω is a closed interval, there are several versions of these sorts of embeddings which are enormously useful in the study of nonlinear evolution equations or inclusions.

The following theorem is an infinite dimensional version of the Ascoli Arzela theorem. It is like a well known result due to Simon [51]. It is an appropriate generalization when you do not necessarily have weak derivatives. I am giving another proof although Theorem 24.11.3 given above is actually more general.

Theorem 24.11.6 *Let $q > 1$ and let $E \subseteq W \subseteq X$ where the injection map is continuous from W to X and compact from E to W. Let S be defined by*

$$\left\{u \text{ such that } \|u(t)\|_E \leq R \text{ for all } t \in [a,b], \text{ and } \|u(s) - u(t)\|_X \leq R|t-s|^{1/q}\right\}.$$

Thus S is bounded in $L^\infty(a,b,E)$ and in addition, the functions are uniformly Holder continuous into X. Then $S \subseteq C([a,b]; W)$ and if $\{u_n\} \subseteq S$, there exists a subsequence, $\{u_{n_k}\}$ which converges to a function $u \in C([a,b]; W)$ in the following way.

$$\lim_{k\to\infty} \|u_{n_k} - u\|_{\infty, W} = 0.$$

Proof: First consider the issue of S being a subset of $C([a,b]; W)$. Let $\varepsilon > 0$ be given. Then by Lemma 24.11.1, there exists a constant, C_ε such that for all $u \in W$

$$\|u\|_W \leq \frac{\varepsilon}{6R} \|u\|_E + C_\varepsilon \|u\|_X.$$

Therefore, for all $u \in S$,

$$\begin{aligned} \|u(t)-u(s)\|_W &\leq \frac{\varepsilon}{6R}\|u(t)-u(s)\|_E + C_\varepsilon \|u(t)-u(s)\|_X \\ &\leq \frac{\varepsilon}{6R}(\|u(t)\|_E + \|u(s)\|_E) + C_\varepsilon \|u(t)-u(s)\|_X \\ &\leq \frac{\varepsilon}{3} + C_\varepsilon R|t-s|^{1/q}. \end{aligned} \tag{24.47}$$

Since ε is arbitrary, it follows $u \in C([a,b];W)$.

Let $D = \mathbb{Q} \cap [a,b]$ so D is a countable dense subset of $[a,b]$. Let $D = \{t_n\}_{n=1}^{\infty}$. By compactness of the embedding of E into W, there exists a subsequence $u_{(n,1)}$ such that as $n \to \infty$, $u_{(n,1)}(t_1)$ converges to a point in W. Now take a subsequence of this, called $(n,2)$ such that as $n \to \infty, u_{(n,2)}(t_2)$ converges to a point in W. It follows that $u_{(n,2)}(t_1)$ also converges to a point of W. Continue this way. Now consider the diagonal sequence, $u_k \equiv u_{(k,k)}$ This sequence is a subsequence of $u_{(n,l)}$ whenever $k > l$. Therefore, $u_k(t_j)$ converges for all $t_j \in D$.

Claim: Let $\{u_k\}$ be as just defined, converging at every point of $D \equiv [a,b] \cap \mathbb{Q}$. Then $\{u_k\}$ converges at every point of $[a,b]$.

Proof of claim: Let $\varepsilon > 0$ be given. Let $t \in [a,b]$. Pick $t_m \in D \cap [a,b]$ such that in 24.47 $C_\varepsilon R|t-t_m| < \varepsilon/3$. Then there exists N such that if $l,n > N$, then $||u_l(t_m) - u_n(t_m)||_X < \varepsilon/3$. It follows that for $l,n > N$,

$$\begin{aligned} \|u_l(t)-u_n(t)\|_W &\leq \|u_l(t)-u_l(t_m)\|_W + \|u_l(t_m)-u_n(t_m)\|_W \\ &\quad + \|u_n(t_m)-u_n(t)\|_W \\ &\leq \frac{2\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{2\varepsilon}{3} < 2\varepsilon \end{aligned}$$

Since ε was arbitrary, this shows $\{u_k(t)\}_{k=1}^{\infty}$ is a Cauchy sequence. Since W is complete, this shows this sequence converges.

Now for $t \in [a,b]$, it was just shown that if $\varepsilon > 0$ there exists N_t such that if $n,m > N_t$, then

$$\|u_n(t)-u_m(t)\|_W < \frac{\varepsilon}{3}.$$

Now let $s \neq t$. Then

$$\|u_n(s)-u_m(s)\|_W \leq \|u_n(s)-u_n(t)\|_W + \|u_n(t)-u_m(t)\|_W + \|u_m(t)-u_m(s)\|_W$$

From 24.47

$$\|u_n(s)-u_m(s)\|_W \leq 2\left(\frac{\varepsilon}{3} + C_\varepsilon R|t-s|^{1/q}\right) + \|u_n(t)-u_m(t)\|_W$$

and so it follows that if δ is sufficiently small and $s \in B(t,\delta)$, then when $n,m > N_t$

$$\|u_n(s)-u_m(s)\| < \varepsilon.$$

Since $[a,b]$ is compact, there are finitely many of these balls, $\{B(t_i,\delta)\}_{i=1}^{p}$, such that for $s \in B(t_i,\delta)$ and $n,m > N_{t_i}$, the above inequality holds. Let $N > \max\{N_{t_1},\cdots,N_{t_p}\}$. Then if $m,n > N$ and $s \in [a,b]$ is arbitrary, it follows the above inequality must hold. Therefore, this has shown the following claim.

Claim: Let $\varepsilon > 0$ be given. Then there exists N such that if $m,n > N$, then it follows that $||u_n - u_m||_{\infty,W} < \varepsilon$.

Now let $u(t) = \lim_{k\to\infty} u_k(t)$.

$$\|u(t) - u(s)\|_W \le \|u(t) - u_n(t)\|_W + \|u_n(t) - u_n(s)\|_W + \|u_n(s) - u(s)\|_W \tag{24.48}$$

Let N be in the above claim and fix $n > N$. Then

$$\|u(t) - u_n(t)\|_W = \lim_{m\to\infty} \|u_m(t) - u_n(t)\|_W \le \varepsilon$$

and similarly, $\|u_n(s) - u(s)\|_W \le \varepsilon$. Then if $|t-s|$ is small enough, 24.47 shows the middle term in 24.48 is also smaller than ε. Therefore, if $|t-s|$ is small enough,

$$\|u(t) - u(s)\|_W < 3\varepsilon.$$

Thus u is continuous. Finally, let N be as in the above claim. Then letting $m,n > N$, it follows that for all $t \in [a,b]$,

$$\|u_m(t) - u_n(t)\|_W < \varepsilon.$$

Therefore, letting $m \to \infty$, it follows that for all $t \in [a,b]$,

$$\|u(t) - u_n(t)\|_W \le \varepsilon.$$

and so $\|u - u_n\|_{\infty,W} \le \varepsilon$. ■

Here is an interesting corollary. Recall that for E a Banach space $C^{0,\alpha}([0,T],E)$ is the space of continuous functions u from $[0,T]$ to E such that

$$\|u\|_{\alpha,E} \equiv \|u\|_{\infty,E} + \rho_{\alpha,E}(u) < \infty$$

where here

$$\rho_{\alpha,E}(u) \equiv \sup_{t\neq s} \frac{\|u(t) - u(s)\|_E}{|t-s|^\alpha}$$

Corollary 24.11.7 *Let $E \subseteq W \subseteq X$ where the injection map is continuous from W to X and compact from E to W. Then if $\gamma > \alpha$, the embedding of $C^{0,\gamma}([0,T],E)$ into $C^{0,\alpha}([0,T],X)$ is compact.*

Proof: Let $\phi \in C^{0,\gamma}([0,T],E)$

$$\frac{\|\phi(t) - \phi(s)\|_X}{|t-s|^\alpha} \le \left(\frac{\|\phi(t) - \phi(s)\|_W}{|t-s|^\gamma}\right)^{\alpha/\gamma} \|\phi(t) - \phi(s)\|_W^{1-(\alpha/\gamma)}$$

$$\le \left(\frac{\|\phi(t) - \phi(s)\|_E}{|t-s|^\gamma}\right)^{\alpha/\gamma} \|\phi(t) - \phi(s)\|_W^{1-(\alpha/\gamma)} \le \rho_{\gamma,E}(\phi)\, \|\phi(t) - \phi(s)\|_W^{1-(\alpha/\gamma)}$$

Now suppose $\{u_n\}$ is a bounded sequence in $C^{0,\gamma}([0,T],E)$. By Theorem 24.11.6 above, there is a subsequence still called $\{u_n\}$ which converges in $C^0([0,T],W)$. Thus from the above inequality

$$\begin{aligned} & \frac{\|u_n(t) - u_m(t) - (u_n(s) - u_m(s))\|_X}{|t-s|^\alpha} \\ \le\ & \rho_{\gamma,E}(u_n - u_m)\, \|u_n(t) - u_m(t) - (u_n(s) - u_m(s))\|_W^{1-(\alpha/\gamma)} \\ \le\ & C(\{u_n\}) \left(2\|u_n - u_m\|_{\infty,W}\right)^{1-(\alpha/\gamma)} \end{aligned}$$

which converges to 0 as $n,m \to \infty$. Thus, $\rho_{\alpha,X}(u_n - u_m) \to 0$ as $n,m \to \infty$ Also

$$\|u_n - u_m\|_{\infty,X} \to 0$$

as $n,m \to \infty$ so this sequence is a Cauchy sequence in $C^{0,\alpha}([0,T],X)$. ■

The next theorem is a well known result probably due to Lions, Teman, or Aubin.

Theorem 24.11.8 *Let $E \subseteq W \subseteq X$ where the injection map is continuous from W to X and compact from E to W. Let $p \geq 1$, let $q > 1$, and define*

$$S \equiv \{u \in L^p([a,b];E) : \text{for some } C,\ \|u(t) - u(s)\|_X \leq C|t-s|^{1/q}$$

$$\text{and } \|u\|_{L^p([a,b];E)} \leq R\}.$$

Thus S is bounded in $L^p([a,b];E)$ and Holder continuous into X. Then S is precompact in $L^p([a,b];W)$. This means that if $\{u_n\}_{n=1}^{\infty} \subseteq S$, it has a subsequence $\{u_{n_k}\}$ which converges in $L^p([a,b];W)$.

Proof: It suffices to show S has an η net in $L^p([a,b];W)$ for each $\eta > 0$.

If not, there exists $\eta > 0$ and a sequence $\{u_n\} \subseteq S$, such that

$$\|u_n - u_m\| \geq \eta \tag{24.49}$$

for all $n \neq m$ and the norm refers to $L^p([a,b];W)$. Let

$$a = t_0 < t_1 < \cdots < t_k = b,\ t_i - t_{i-1} = (b-a)/k.$$

Now define

$$\overline{u}_n(t) \equiv \sum_{i=1}^{k} \overline{u}_{n_i} \mathcal{X}_{[t_{i-1},t_i)}(t),\ \overline{u}_{n_i} \equiv \frac{1}{t_i - t_{i-1}} \int_{t_{i-1}}^{t_i} u_n(s)\,ds.$$

The idea is to show that $\overline{u}_n$ approximates u_n well and then to argue that a subsequence of the $\{\overline{u}_n\}$ is a Cauchy sequence yielding a contradiction to 24.49.

Therefore,

$$u_n(t) - \overline{u}_n(t) = \sum_{i=1}^{k} u_n(t)\,\mathcal{X}_{[t_{i-1},t_i)}(t) - \sum_{i=1}^{k} \overline{u}_{n_i} \mathcal{X}_{[t_{i-1},t_i)}(t)$$

$$= \sum_{i=1}^{k} \frac{1}{t_i - t_{i-1}} \int_{t_{i-1}}^{t_i} u_n(t)\,ds \mathcal{X}_{[t_{i-1},t_i)}(t) - \sum_{i=1}^{k} \frac{1}{t_i - t_{i-1}} \int_{t_{i-1}}^{t_i} u_n(s)\,ds \mathcal{X}_{[t_{i-1},t_i)}(t)$$

$$= \sum_{i=1}^{k} \frac{1}{t_i - t_{i-1}} \int_{t_{i-1}}^{t_i} (u_n(t) - u_n(s))\,ds \mathcal{X}_{[t_{i-1},t_i)}(t).$$

It follows from Jensen's inequality that

$$\begin{aligned} & ||u_n(t) - \overline{u}_n(t)||_W^p \\ = & \sum_{i=1}^{k} \left\| \frac{1}{t_i - t_{i-1}} \int_{t_{i-1}}^{t_i} (u_n(t) - u_n(s))\,ds \right\|_W^p \mathcal{X}_{[t_{i-1},t_i)}(t) \\ \leq & \sum_{i=1}^{k} \frac{1}{t_i - t_{i-1}} \int_{t_{i-1}}^{t_i} \|u_n(t) - u_n(s)\|_W^p\,ds \mathcal{X}_{[t_{i-1},t_i)}(t) \end{aligned}$$

and so

$$\begin{aligned}
&\int_a^b \|u_n(t)-\overline{u}_n(s)\|_W^p\,ds \\
\leq\ &\int_a^b \sum_{i=1}^k \frac{1}{t_i-t_{i-1}}\int_{t_{i-1}}^{t_i}\|u_n(t)-u_n(s)\|_W^p\,ds\,\mathcal{X}_{[t_{i-1},t_i)}(t)\,dt \\
=\ &\sum_{i=1}^k \frac{1}{t_i-t_{i-1}}\int_{t_{i-1}}^{t_i}\int_{t_{i-1}}^{t_i}\|u_n(t)-u_n(s)\|_W^p\,dsdt. \qquad (24.50)
\end{aligned}$$

From Lemma 24.11.1 if $\varepsilon>0$, there exists C_ε such that

$$\begin{aligned}
\|u_n(t)-u_n(s)\|_W^p &\leq \varepsilon\|u_n(t)-u_n(s)\|_E^p + C_\varepsilon\|u_n(t)-u_n(s)\|_X^p \\
&\leq 2^{p-1}\varepsilon\left(\|u_n(t)\|^p+\|u_n(s)\|^p\right)+C_\varepsilon|t-s|^{p/q}
\end{aligned}$$

This is substituted in to 24.50 to obtain

$$\int_a^b \|u_n(t)-\overline{u}_n(s)\|_W^p\,ds \leq$$

$$\begin{aligned}
&\sum_{i=1}^k \frac{1}{t_i-t_{i-1}}\int_{t_{i-1}}^{t_i}\int_{t_{i-1}}^{t_i}\left(2^{p-1}\varepsilon\left(\|u_n(t)\|^p+\|u_n(s)\|^p\right)+C_\varepsilon|t-s|^{p/q}\right)dsdt \\
=\ &\sum_{i=1}^k 2^p\varepsilon\int_{t_{i-1}}^{t_i}\|u_n(t)\|_W^p + \frac{C_\varepsilon}{t_i-t_{i-1}}\int_{t_{i-1}}^{t_i}\int_{t_{i-1}}^{t_i}|t-s|^{p/q}\,dsdt \\
\leq\ &2^p\varepsilon\int_a^b\|u_n(t)\|^p\,dt + C_\varepsilon\sum_{i=1}^k\frac{1}{(t_i-t_{i-1})}(t_i-t_{i-1})^{p/q}\int_{t_{i-1}}^{t_i}\int_{t_{i-1}}^{t_i}dsdt \\
=\ &2^p\varepsilon\int_a^b\|u_n(t)\|^p\,dt + C_\varepsilon\sum_{i=1}^k\frac{1}{(t_i-t_{i-1})}(t_i-t_{i-1})^{p/q}(t_i-t_{i-1})^2 \\
\leq\ &2^p\varepsilon R^p + C_\varepsilon\sum_{i=1}^k(t_i-t_{i-1})^{1+p/q} = 2^p\varepsilon R^p + C_\varepsilon k\left(\frac{b-a}{k}\right)^{1+p/q}.
\end{aligned}$$

Taking ε so small that $2^p\varepsilon R^p<\eta^p/8^p$ and then choosing k sufficiently large, it follows

$$||u_n-\overline{u}_n||_{L^p([a,b];W)}<\frac{\eta}{4}.$$

Thus k is fixed and $\overline{u}_n$ at a step function with k steps having values in E. Now use compactness of the embedding of E into W to obtain a subsequence such that $\{\overline{u}_n\}$ is Cauchy in $L^p(a,b;W)$ and use this to contradict 24.49. The details follow.

Suppose $\overline{u}_n(t)=\sum_{i=1}^k u_i^n\mathcal{X}_{[t_{i-1},t_i)}(t)$. Thus

$$\|\overline{u}_n(t)\|_E=\sum_{i=1}^k\|u_i^n\|_E\,\mathcal{X}_{[t_{i-1},t_i)}(t)$$

and so

$$R\geq\int_a^b\|\overline{u}_n(t)\|_E^p\,dt=\frac{T}{k}\sum_{i=1}^k\|u_i^n\|_E^p$$

Therefore, the $\{u_i^n\}$ are all bounded. It follows that after taking subsequences k times there exists a subsequence $\{u_{n_k}\}$ such that u_{n_k} is a Cauchy sequence in $L^p(a,b;W)$. You simply get a subsequence such that $u_i^{n_k}$ is a Cauchy sequence in W for each i. Then denoting this subsequence by n,

$$\begin{aligned}\|u_n-u_m\|_{L^p(a,b;W)} &\leq \|u_n-\overline{u}_n\|_{L^p(a,b;W)} \\ &\quad +\|\overline{u}_n-\overline{u}_m\|_{L^p(a,b;W)}+\|\overline{u}_m-u_m\|_{L^p(a,b;W)} \\ &\leq \frac{\eta}{4}+\|\overline{u}_n-\overline{u}_m\|_{L^p(a,b;W)}+\frac{\eta}{4}<\eta\end{aligned}$$

provided m,n are large enough, contradicting 24.49. ■

You can give a different version of the above to include the case where there is, instead of a Holder condition, a bound on u' for $u\in S$. It is stated next. We are assuming a situation in which

$$\int_a^b u'(t)\,dt = u(b)-u(a)$$

This happens, for example, if u' is the weak derivative. This is discussed in the exercises. These kind of theorems are in [51].

Corollary 24.11.9 *Let $E\subseteq W\subseteq X$ where the injection map is continuous from W to X and compact from E to W. Let $p\geq 1$, let $q>1$, and define*

$$S\equiv\{u\in L^p([a,b];E) : \text{for some } C,\ \|u(t)-u(s)\|_X\leq C|t-s|^{1/q}$$

$$\text{and } \|u\|_{L^p([a,b];E)}\leq R\}.$$

Thus S is bounded in $L^p([a,b];E)$ and Holder continuous into X. Then S is precompact in $L^p([a,b];W)$. This means that if $\{u_n\}_{n=1}^{\infty}\subseteq S$, it has a subsequence $\{u_{n_k}\}$ which converges in $L^p([a,b];W)$. The same conclusion can be drawn if it is known instead of the Holder condition that $\|u'\|_{L^1([a,b];X)}$ is bounded.

Proof: The first part is Theorem 24.11.8. Therefore, we just prove the new stuff which involves a bound on the L^1 norm of the derivative. It suffices to show S has an η net in $L^p([a,b];W)$ for each $\eta>0$.

If not, there exists $\eta>0$ and a sequence $\{u_n\}\subseteq S$, such that

$$\|u_n-u_m\|\geq\eta \tag{24.51}$$

for all $n\neq m$ and the norm refers to $L^p([a,b];W)$. Let

$$a=t_0<t_1<\cdots<t_k=b,\ t_i-t_{i-1}=(b-a)/k.$$

Now define

$$\overline{u}_n(t)\equiv\sum_{i=1}^{k}\overline{u}_{n_i}\mathcal{X}_{[t_{i-1},t_i)}(t),\ \overline{u}_{n_i}\equiv\frac{1}{t_i-t_{i-1}}\int_{t_{i-1}}^{t_i}u_n(s)\,ds.$$

The idea is to show that $\overline{u}_n$ approximates u_n well and then to argue that a subsequence of the $\{\overline{u}_n\}$ is a Cauchy sequence yielding a contradiction to 24.51.

Therefore,

$$u_n(t)-\overline{u}_n(t)=\sum_{i=1}^{k}u_n(t)\,\mathcal{X}_{[t_{i-1},t_i)}(t)-\sum_{i=1}^{k}\overline{u}_{n_i}\mathcal{X}_{[t_{i-1},t_i)}(t)$$

$$= \sum_{i=1}^{k} \frac{1}{t_i - t_{i-1}} \int_{t_{i-1}}^{t_i} u_n(t)\, ds \mathscr{X}_{[t_{i-1},t_i)}(t) - \sum_{i=1}^{k} \frac{1}{t_i - t_{i-1}} \int_{t_{i-1}}^{t_i} u_n(s)\, ds \mathscr{X}_{[t_{i-1},t_i)}(t)$$

$$= \sum_{i=1}^{k} \frac{1}{t_i - t_{i-1}} \int_{t_{i-1}}^{t_i} (u_n(t) - u_n(s))\, ds \mathscr{X}_{[t_{i-1},t_i)}(t).$$

It follows from Jensen's inequality that

$$\begin{aligned} & \|u_n(t) - \overline{u}_n(t)\|_W^p \\ = & \sum_{i=1}^{k} \left\| \frac{1}{t_i - t_{i-1}} \int_{t_{i-1}}^{t_i} (u_n(t) - u_n(s))\, ds \right\|_W^p \mathscr{X}_{[t_{i-1},t_i)}(t) \end{aligned}$$

And so

$$\int_0^T \|u_n(t) - \overline{u}_n(t)\|_W^p\, dt = \sum_{i=1}^{k} \int_{t_{i-1}}^{t_i} \left\| \frac{1}{t_i - t_{i-1}} \int_{t_{i-1}}^{t_i} (u_n(t) - u_n(s))\, ds \right\|_W^p dt$$

$$\begin{aligned} \leq & \sum_{i=1}^{k} \int_{t_{i-1}}^{t_i} \varepsilon \left\| \frac{1}{t_i - t_{i-1}} \int_{t_{i-1}}^{t_i} (u_n(t) - u_n(s))\, ds \right\|_E^p dt \\ & + C_\varepsilon \sum_{i=1}^{k} \int_{t_{i-1}}^{t_i} \left\| \frac{1}{t_i - t_{i-1}} \int_{t_{i-1}}^{t_i} (u_n(t) - u_n(s))\, ds \right\|_X^p dt \end{aligned} \tag{24.52}$$

Consider the second of these. It equals

$$C_\varepsilon \sum_{i=1}^{k} \int_{t_{i-1}}^{t_i} \left\| \frac{1}{t_i - t_{i-1}} \int_{t_{i-1}}^{t_i} \int_s^t u_n'(\tau)\, d\tau ds \right\|_X^p dt$$

This is no larger than

$$\leq C_\varepsilon \sum_{i=1}^{k} \int_{t_{i-1}}^{t_i} \left(\frac{1}{t_i - t_{i-1}} \int_{t_{i-1}}^{t_i} \int_{t_{i-1}}^{t_i} \|u_n'(\tau)\|_X\, d\tau ds \right)^p dt$$

$$= C_\varepsilon \sum_{i=1}^{k} \int_{t_{i-1}}^{t_i} \left(\int_{t_{i-1}}^{t_i} \|u_n'(\tau)\|_X\, d\tau \right)^p dt$$

$$= C_\varepsilon \sum_{i=1}^{k} \left((t_i - t_{i-1})^{1/p} \int_{t_{i-1}}^{t_i} \|u_n'(\tau)\|_X\, d\tau \right)^p$$

Since $\frac{b-a}{k} = t_i - t_{i-1}$,

$$\begin{aligned} = & C_\varepsilon \left(\sum_{i=1}^{k} (t_i - t_{i-1})^{1/p} \int_{t_{i-1}}^{t_i} \|u_n'(\tau)\|_X\, d\tau \right)^p \\ = & C_\varepsilon \left(\sum_{i=1}^{k} \left(\frac{b-a}{k} \right)^{1/p} \int_{t_{i-1}}^{t_i} \|u_n'(\tau)\|_X\, d\tau \right)^p \\ \leq & \frac{C_\varepsilon (b-a)}{k} \left(\sum_{i=1}^{k} \int_{t_{i-1}}^{t_i} \|u_n'(\tau)\|_X\, d\tau \right)^p \\ = & \frac{C_\varepsilon (b-a)}{k} \left(\|u_n'\|_{L^1([a,b],X)} \right)^p < \frac{\eta^p}{10^p} \end{aligned}$$

if k is chosen large enough. Now consider the first in 24.52. By Jensen's inequality

$$\sum_{i=1}^{k}\int_{t_{i-1}}^{t_i}\varepsilon\left\|\frac{1}{t_i-t_{i-1}}\int_{t_{i-1}}^{t_i}\left(u_n(t)-u_n(s)\right)ds\right\|_E^p dt \leq$$

$$\sum_{i=1}^{k}\int_{t_{i-1}}^{t_i}\varepsilon\frac{1}{t_i-t_{i-1}}\int_{t_{i-1}}^{t_i}\|u_n(t)-u_n(s)\|_E^p\,dsdt$$

$$\begin{aligned}\leq\ & \varepsilon 2^{p-1}\sum_{i=1}^{k}\frac{1}{t_i-t_{i-1}}\int_{t_{i-1}}^{t_i}\int_{t_{i-1}}^{t_i}\left(\|u_n(t)\|^p+\|u_n(s)\|^p\right)dsdt\\ =\ & 2\varepsilon 2^{p-1}\sum_{i=1}^{k}\int_{t_{i-1}}^{t_i}\left(\|u_n(t)\|^p\right)dt=\varepsilon(2)\left(2^{p-1}\right)\|u_n\|_{L^p([a,b],E)}\leq M\varepsilon\end{aligned}$$

Now pick ε sufficiently small that $M\varepsilon<\frac{\eta^p}{10^p}$ and then k large enough that the second term in 24.52 is also less than $\eta^p/10^p$. Then it will follow that

$$\|\bar{u}_n-u_n\|_{L^p([a,b],W)}<\left(\frac{2\eta^p}{10^p}\right)^{1/p}=2^{1/p}\frac{\eta}{10}\leq\frac{\eta}{5}$$

Thus k is fixed and $\bar{u}_n$ at a step function with k steps having values in E. Now use compactness of the embedding of E into W to obtain a subsequence such that $\{\bar{u}_n\}$ is Cauchy in $L^p([a,b];W)$ and use this to contradict 24.51. The details follow.

Suppose $\bar{u}_n(t)=\sum_{i=1}^{k}u_i^n\mathcal{X}_{[t_{i-1},t_i)}(t)$. Thus

$$\|\bar{u}_n(t)\|_E=\sum_{i=1}^{k}\|u_i^n\|_E\,\mathcal{X}_{[t_{i-1},t_i)}(t)$$

and so

$$R\geq\int_a^b\|\bar{u}_n(t)\|_E^p\,dt=\frac{T}{k}\sum_{i=1}^{k}\|u_i^n\|_E^p$$

Therefore, the $\{u_i^n\}$ are all bounded. It follows that after taking subsequences k times there exists a subsequence $\{u_{n_k}\}$ such that u_{n_k} is a Cauchy sequence in $L^p([a,b];W)$. You simply get a subsequence such that $u_i^{n_k}$ is a Cauchy sequence in W for each i. Then denoting this subsequence by n,

$$\begin{aligned}\|u_n-u_m\|_{L^p(a,b;W)}\ &\leq\ \|u_n-\bar{u}_n\|_{L^p(a,b;W)}\\ &\quad+\|\bar{u}_n-\bar{u}_m\|_{L^p(a,b;W)}+\|\bar{u}_m-u_m\|_{L^p(a,b;W)}\\ &\leq\ \frac{\eta}{4}+\|\bar{u}_n-\bar{u}_m\|_{L^p(a,b;W)}+\frac{\eta}{4}<\eta\end{aligned}$$

provided m,n are large enough, contradicting 24.51. ■

24.12 Conditional Expectation in Banach Spaces

Let $(\Omega,\mathcal{F},P)$ be a probability space and let $X\in L^1(\Omega;\mathbb{R})$. Also let $\mathcal{G}\subseteq\mathcal{F}$ where $\mathcal{G}$ is also a σ algebra. Then the usual conditional expectation is defined by

$$\int_A XdP=\int_A E(X|\mathcal{G})\,dP$$

where $E(X|\mathscr{G})$ is $\mathscr{G}$ measurable and $A \in \mathscr{G}$ is arbitrary. Recall this is an application of the Radon Nikodym theorem. Also recall $E(X|\mathscr{G})$ is unique up to a set of measure zero.

I want to do something like this here. Denote by $L^1(\Omega;E,\mathscr{G})$ those functions in $L^1(\Omega;E)$ which are measurable with respect to $\mathscr{G}$.

Theorem 24.12.1 *Let E be a separable Banach space and let*

$$X \in L^1(\Omega;E,\mathscr{F})$$

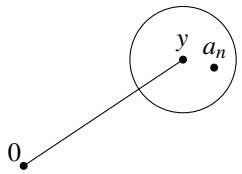
where X is measurable with respect to $\mathscr{F}$ and let $\mathscr{G}$ be a σ algebra which is contained in $\mathscr{F}$. Then there exists a unique $Z \in L^1(\Omega;E,\mathscr{G})$ such that for all $A \in \mathscr{G}$,

$$\int_A X dP = \int_A Z dP$$

Denoting this Z as $E(X|\mathscr{G})$, it follows

$$\|E(X|\mathscr{G})\| \leq E(\|X\| \ |\mathscr{G}).$$

Proof: First consider uniqueness. Suppose Z' is another in $L^1(\Omega;E,\mathscr{G})$ which works. Consider a dense subset of E $\{a_n\}_{n=1}^{\infty}$. Then the balls $\left\{B\left(a_n, \frac{\|a_n\|}{4}\right)\right\}_{n=1}^{\infty}$ must cover $E \setminus \{\mathbf{0}\}$. Here is why. If $y \neq 0$, pick $a_n \in B\left(y, \frac{\|y\|}{5}\right)$.



Then $\|a_n\| \geq 4\|y\|/5$ and so $\|a_n - y\| < \|y\|/5$. Thus

$$y \in B(a_n, \|y\|/5) \subseteq B\left(a_n, \frac{\|a_n\|}{4}\right)$$

Now suppose Z is $\mathscr{G}$ measurable and

$$\int_A Z dP = 0$$

for all $A \in \mathscr{G}$. The letting $A \equiv Z^{-1}\left(B\left(a_n, \frac{\|a_n\|}{4}\right)\right)$ it follows

$$0 = \int_A Z - a_n + a_n dP$$

and so

$$\|a_n\| P(A) = \left\| \int_A a_n dP \right\| = \left\| \int_A (a_n - Z) dP \right\|$$

$$\leq \int_{Z^{-1}\left(B\left(a_n, \frac{\|a_n\|}{4}\right)\right)} \|a_n - Z\| dP \leq \frac{\|a_n\|}{4} P(A)$$

which is a contradiction unless $P(A) = 0$. Therefore, letting

$$N \equiv \cup_{n=1}^{\infty} Z^{-1}\left(B\left(a_n, \frac{\|a_n\|}{4}\right)\right) = Z^{-1}(E \setminus \{0\})$$

it follows N has measure zero and so $Z = 0$ a.e. This proves uniqueness because if Z, Z' both hold, then from the above argument, $Z - Z' = 0$ a.e.

Next I will show Z exists. To do this recall Theorem 24.2.4 on Page 679 which is stated below for convenience.

Theorem 24.12.2 *An E valued function, X, is Bochner integrable if and only if X is strongly measurable and*

$$\int_{\Omega} \|X(\omega)\| \, dP < \infty. \tag{24.53}$$

In this case there exists a sequence of simple functions $\{X_n\}$ satisfying

$$\int_{\Omega} \|X_n(\omega) - X_m(\omega)\| \, dP \to 0 \text{ as } m, n \to \infty. \tag{24.54}$$

$X_n(\omega)$ converging pointwise to $X(\omega)$,

$$\|X_n(\omega)\| \leq 2\|X(\omega)\| \tag{24.55}$$

and

$$\lim_{n\to\infty} \int_{\Omega} \|X(\omega) - X_n(\omega)\| \, dP = 0. \tag{24.56}$$

Now let $\{X_n\}$ be the simple functions just defined and let

$$X_n(\omega) = \sum_{k=1}^{m} x_k \mathcal{X}_{F_k}(\omega)$$

where $F_k \in \mathcal{F}$, the F_k being disjoint. Then define

$$Z_n \equiv \sum_{k=1}^{m} x_k E\left(\mathcal{X}_{F_k} | \mathcal{G}\right).$$

Thus, if $A \in \mathcal{G}$,

$$\begin{aligned} \int_A Z_n dP &= \sum_{k=1}^{m} x_k \int_A E\left(\mathcal{X}_{F_k} | \mathcal{G}\right) dP \\ &= \sum_{k=1}^{m} x_k \int_A \mathcal{X}_{F_k} dP \\ &= \sum_{k=1}^{m} x_k P(F_k \cap A) = \int_A X_n dP \end{aligned} \tag{24.57}$$

Then since $E\left(\mathcal{X}_{F_k} | \mathcal{G}\right) \geq 0$,

$$\|Z_n\| \leq \sum_{k=1}^{m} \|x_k\| E\left(\mathcal{X}_{F_k} | \mathcal{G}\right)$$

Thus if $A \in \mathcal{G}$,

$$\begin{aligned} E(\|Z_n\| \mathcal{X}_A) &\leq E\left(\sum_{k=1}^{m} \|x_k\| \mathcal{X}_A E\left(\mathcal{X}_{F_k}|\mathcal{G}\right)\right) = \sum_{k=1}^{m} \|x_k\| \int_A E\left(\mathcal{X}_{F_k}|\mathcal{G}\right) dP \\ &= \sum_{k=1}^{m} \|x_k\| \int_A \mathcal{X}_{F_k} dP = E(\mathcal{X}_A \|X_n\|). \end{aligned} \tag{24.58}$$

Note the use of $\leq$ in the first step in the above. Although the F_k are disjoint, all that is known about $E\left(\mathcal{X}_{F_k}|\mathcal{G}\right)$ is that it is nonnegative. Similarly,

$$E(\|Z_n - Z_m\|) \leq E(\|X_n - X_m\|)$$

and this last term converges to 0 as $n, m \to \infty$ by the properties of the X_n. Therefore, $\{Z_n\}$ is a Cauchy sequence in $L^1(\Omega; E; \mathcal{G})$. It follows it converges to some Z in $L^1(\Omega; E, \mathcal{G})$. Then letting $A \in \mathcal{G}$, and using 24.57,

$$\begin{aligned} \int_A Z dP &= \int \mathcal{X}_A Z dP = \lim_{n\to\infty} \int \mathcal{X}_A Z_n dP = \lim_{n\to\infty} \int_A Z_n dP \\ &= \lim_{n\to\infty} \int_A X_n dP = \int_A X dP. \end{aligned}$$

Then define $Z \equiv E(X|\mathcal{G})$.

It remains to verify $\|E(X|\mathcal{G})\| \equiv \|Z\| \leq E(\|X\| \ |\mathcal{G})$. This follows because, from the above,

$$\|Z_n\| \to \|Z\|, \ \|X_n\| \to \|X\| \text{ in } L^1(\Omega)$$

and so if $A \in \mathcal{G}$, then from 24.58,

$$\frac{1}{P(A)} \int_A \|Z_n\| dP \leq \frac{1}{P(A)} \int_A \|X_n\| dP$$

and so, passing to the limit,

$$\frac{1}{P(A)} \int_A \|Z\| dP \leq \frac{1}{P(A)} \int_A \|X\| dP = \frac{1}{P(A)} \int_A E(\|X\| \ |\mathcal{G}) dP$$

Since A is arbitrary, this shows that

$$\|E(X|\mathcal{G})\| \equiv \|Z\| \leq E(\|X\| \ |\mathcal{G}). \blacksquare$$

In the case where E is reflexive, one could also use Corollary 24.7.6 on Page 706 to get the above result. You would define a vector measure on $\mathcal{G}$,

$$\nu(F) \equiv \int_F X dP$$

and then you would use the fact that reflexive separable Banach spaces have the Radon Nikodym property to obtain $Z \in L^1(\Omega; E, \mathcal{G})$ such that

$$\nu(F) = \int_F X dP = \int_F Z dP.$$

The function, Z whose existence and uniqueness is guaranteed by Theorem 24.12.2 is called $E(X|\mathcal{G})$.

24.13 Exercises

1. Show $L^1(\mathbb{R})$ is not reflexive. **Hint:** $L^1(\mathbb{R})$ is separable. What about $L^\infty(\mathbb{R})$?

2. If $f \in L^1(\mathbb{R}^n;X)$ for X a Banach space, does the usual fundamental theorem of calculus work? That is, can you say $\lim_{r\to 0} \frac{1}{m(B(\boldsymbol{x},r))} \int_{B(\boldsymbol{x},r)} f(t)\,dm = f(x)$ a.e.?

3. Does the Vitali convergence theorem hold for Bochner integrable functions? If so, give a statement of the appropriate theorem and a proof.

4. Suppose $g \in L^1([a,b];X)$ where X is a Banach space. Then if $\int_a^b g(t)\,\phi(t)\,dt = 0$ for all $\phi \in C_c^\infty(a,b)$, then $g(t) = 0$ a.e. Show that this is the case. **Hint:** It will likely depend on the regularity properties of Lebesgue measure.

5. Suppose $f \in L^1(a,b;X)$ and for all $\phi \in C_c^\infty(a,b)$, $\int_a^b f(t)\,\phi'(t)\,dt = 0$.Then there exists a constant, $a \in X$ such that $f(t) = a$ a.e. **Hint:** Let

$$\psi_\phi(x) \equiv \int_a^x [\phi(t) - \left(\int_a^b \phi(y)\,dy\right)\phi_0(t)]dt,\ \phi_0 \in C_c^\infty(a,b), \int_a^b \phi_0(x)\,dx = 1$$

   Then explain why $\psi_\phi \in C_c^\infty(a,b)$, $\psi_\phi' = \phi - \left(\int_a^b \phi(y)\,dy\right)\phi_0$. Then use the assumption on ψ_ϕ. Next use the above problem. Verify that

$$f(y) = \int_a^b f(t)\,\phi_0(t)\,dt \text{ a.e. } y$$

6. Let $f \in L^1([a,b],X)$. Then we say that the weak derivative of f is in $L^1([a,b],X)$ if there is a function denoted as $f' \in L^1([a,b],X)$ such that for all $\phi \in C_c^\infty(a,b)$,

$$-\int_a^b f(t)\,\phi'(t)\,dt = \int_a^b f'(t)\,\phi(t)\,dt$$

   Show that this definition is well defined. Next, using the above problems, show that if $f, f' \in L^1([a,b],X)$, it follows that there is a continuous function, denoted by $t \to \hat{f}(t)$ such that $\hat{f}(t) = f(t)$ a.e. t and

$$\hat{f}(t) = \hat{f}(a) + \int_0^t f'(s)\,ds$$

   Thus, unlike the classical definition of the derivative, when a function and its derivative are both in L^1, it has a representative $\hat{f}$ which equals the function a.e. such that $\hat{f}$ can be recovered from its derivative. Recall the well known example of this not working out which is based on the Cantor function of Problem 4 on Page 273. This function had zero derivative a.e. and yet it climbed from 0 to 1 on the unit interval. Thus one could not recover it from integrating its classical derivative. Incidentally, if the function has a derivative everywhere, then you can recover it by taking the generalized Riemann integral of the derivative, although the Lebesgue integral of this derivative might not even be defined. This is in my book on single variable advanced calculus, but this integral is not discussed here.

Chapter 25

Stone's Theorem and Partitions of Unity

This section is devoted to Stone's theorem which says that a metric space is paracompact, defined below. See [41] for this which is where I read it. First is the definition of what is meant by a refinement. A metric space is an example of a topological space and it is the context for what is done below.

Definition 25.0.1 *Let S be a topological space. We say that a collection of sets $\mathfrak{D}$ is a refinement of an open cover $\mathfrak{S}$, if every set of $\mathfrak{D}$ is contained in some set of $\mathfrak{S}$. An open refinement would be one in which all sets are open, with a similar convention holding for the term " closed refinement".*

Definition 25.0.2 *We say that a collection of sets $\mathfrak{D}$, is locally finite if for all $p \in S$, the topological space, there exists V an open set containing p such that V has nonempty intersection with only finitely many sets of $\mathfrak{D}$.*

Definition 25.0.3 *We say S is paracompact if it is Hausdorff and for every open cover $\mathfrak{S}$, there exists an open refinement $\mathfrak{D}$ such that $\mathfrak{D}$ is locally finite and $\mathfrak{D}$ covers S.*

Recall how the union of finitely many closed sets is closed. This can be generalized to a locally finite set of closed sets. Think $\mathbb{N}$ for example. The following implies this.

Theorem 25.0.4 *If $\mathfrak{D}$ is locally finite then*

$$\cup\{\overline{D} : D \in \mathfrak{D}\} = \overline{\cup\{D : D \in \mathfrak{D}\}}.$$

Proof: It is clear the left side is a subset of the right because the right is a closed set which contains the left since a limit point of any D is in the set on the right . If $p \in \cup\{D : D \in \mathfrak{D}\}$, there is nothing to show. Let p not be in this set but be a limit point of $\cup\{D : D \in \mathfrak{D}\}$. Is p in some $\overline{D}$? Let $p \in V$, an open set intersecting only finitely many sets of $\mathfrak{D}$, $D_1, ..., D_n$. If p is not in any of $\overline{D_i}$ then $p \in W$ where W is some open set which contains no points of $\cup_{i=1}^{n} D_i$. Then $V \cap W$ contains no points of any set of $\mathfrak{D}$ and this contradicts the assumption that p is a limit point of $\cup\{D : D \in \mathfrak{D}\}$. Thus $p \in \overline{D_i}$ for some i. ■

We say $\mathfrak{S} \subseteq \mathscr{P}(S)$ is countably locally finite if

$$\mathfrak{S} = \cup_{n=1}^{\infty} \mathfrak{S}_n$$

and each $\mathfrak{S}_n$ is locally finite. The following theorem appeared in the 1950's. It will be used to prove Stone's theorem.

Theorem 25.0.5 *Let S be a regular topological space. (If $p \in U$ open, then there exists an open set V such that $p \in \bar{V} \subseteq U$.) The following are equivalent*

1.) Every open covering of S has a refinement that is open, covers S and is ***countably*** *locally finite.*

2.) Every open covering of S has a refinement that is locally finite and covers S. (The sets in refinement maybe not open.)

3.) Every open covering of S has a refinement that is closed, locally finite, and covers S. (Sets in refinement are closed.)

4.) Every open covering of S has a refinement that is open, locally finite, and covers S. (Sets in refinement are open.)

Proof:

1.)⇒ 2.)

Let $\mathcal{S}$ be an open cover of S and let $\mathcal{B}$ be an open countably locally finite refinement

$$\mathcal{B} = \cup_{n=1}^{\infty} \mathcal{B}_n$$

where $\mathcal{B}_n$ is an open refinement of $\mathcal{S}$ and $\mathcal{B}_n$ is locally finite. For $B \in \mathcal{B}_n$, let

$$E_n(B) = B \setminus \bigcup_{k<n} (\cup\{B : B \in \mathcal{B}_k\}) \equiv B \cap \left(\bigcup_{k<n} (\cup\{B : B \in \mathcal{B}_k\}) \right)^C.$$

Thus, in words, $E_n(B)$ consists of points in B which are not in any set from any $\mathcal{B}_k$ for $k < n$.

Claim: $\{E_n(B) : n \in \mathbb{N},\ B \in \mathcal{B}_n\}$ is locally finite.

Proof of the claim: Let $p \in S$. Then $p \in B_0 \in \mathcal{B}_n$ for some n. Let V be open, $p \in V$, and V intersects only finitely many sets of $\mathcal{B}_1 \cup ... \cup \mathcal{B}_n$. Then consider $B_0 \cap V$. If $m > n$,

$$(B_0 \cap V) \cap E_m(B) \subseteq \left(\bigcup_{k<m} (\cup\{B : B \in \mathcal{B}_k) \right)^C \subseteq B_0^C.$$

In words, $E_m(B)$ has nothing in it from any of the $\mathcal{B}_k$ for $k < m$. In particular, it has nothing in it from B_0. Thus $(B_0 \cap V) \cap E_m(B) = \emptyset$ for $m > n$. Thus $p \in B_0 \cap V$ which intersects only finitely many sets of $\mathcal{S}$, no more than those intersected by V. This establishes the claim.

Claim: $\{E_n(B) : n \in \mathbb{N},\ B \in \mathcal{B}_n\}$ covers S.

Proof: Let $p \in S$ and let $n = \min\{k \in \mathbb{N} : p \in B \text{ for some } B \in \mathcal{B}_k\}$. Let $p \in B \in \mathcal{B}_n$. Then $p \in E_n(B)$.

The two claims show that 1.)⇒ 2.).

2.)⇒ 3.)

Let $\mathcal{S}$ be an open cover and let

$$\mathcal{G} \equiv \{U : U \text{ is open and } \overline{U} \subseteq V \in \mathcal{S} \text{for some } V \in \mathcal{S}.$$

Then since S is regular, $\mathcal{G}$ covers S. (If $p \in S$, then $p \in U \subseteq \overline{U} \subseteq V \in \mathcal{S}$.) By 2.), $\mathcal{G}$ has a locally finite refinement $\mathcal{C}$, covering S. Consider $\{\overline{E} : E \in \mathcal{C}\}$. This collection of closed sets covers S and is locally finite because if $p \in S$, there exists V, $p \in V$, and V has nonempty intersections with only finitely many elements of $\mathcal{C}$, say $E_1, \cdots, E_n$. If $\overline{E} \cap V \neq \emptyset$, then $E \cap V \neq \emptyset$ and so V intersects only $\overline{E_1}, \cdots, \overline{E_n}$. This shows 2.)⇒ 3.).

3.)⇒ 4.) Here is a table of symbols with a short summary of their meaning.

Open covering	Locally finite refinement
$\mathcal{S}$ original covering	$\mathcal{B}$ by 3. can be closed refinement
$\mathcal{F}$ open intersectors	$\mathcal{C}$ closed refinement

Let $\mathcal{S}$ be an open cover and let $\mathcal{B}$ be a locally finite refinement which covers S. By 3.) we can take $\mathcal{B}$ to be a closed refinement but this is not important here. Let

$$\mathcal{F} \equiv \{U : U \text{ is open and } U \text{ intersects only finitely many sets of } \mathcal{B}\}.$$

Then $\mathfrak{F}$ covers S because $\mathfrak{B}$ is locally finite. (If $p \in S$, then there exists an open set U containing p which intersects only finitely many sets of $\mathfrak{B}$. Thus $p \in U \in \mathfrak{F}$.) By 3., $\mathfrak{F}$ has a locally finite closed refinement $\mathfrak{C}$, which covers S. Define for $B \in \mathfrak{B}$

$$\mathfrak{C}(B) \equiv \{C \in \mathfrak{C} : C \cap B = \emptyset\}$$

Thus these closed sets C do not intersect B and so B is in their complement. We use $\mathfrak{C}(B)$ to fatten up B. Let

$$E(B) \equiv (\cup\{C : C \in \mathfrak{C}(B)\})^C.$$

In words, $E(B)$ is the complement of the union of all closed sets of $\mathfrak{C}$ which do not intersect B. Thus $E(B) \supseteq B$, and has fattened up B. Then since $\mathfrak{C}(B)$ is locally finite, $E(B)$ is an open set by Theorem 25.0.4. Now let $F(B)$ be defined such that for $B \in \mathfrak{B}$,

$$B \subseteq F(B) \in \mathfrak{S}$$

(by definition B is in some set of $\mathfrak{S}$), and let

$$\mathfrak{L} = \{E(B) \cap F(B) : B \in \mathfrak{B}\}$$

The intersection with $F(B)$ is to ensure that $\mathfrak{L}$ is a refinement of $\mathfrak{S}$. The important thing to notice is that **if $C \in \mathfrak{C}$ intersects $E(B)$, then it must also intersect B.** If not, you could include it in the list of closed sets which do not intersect B and whose complement is $E(B)$. Thus $E(B)$ would be too large.

Claim: $\mathfrak{L}$ covers S.

This claim is obvious because if $p \in S$ then $p \in B$ for some $B \in \mathfrak{B}$. Hence

$$p \in E(B) \cap F(B) \in \mathfrak{L}.$$

Claim: $\mathfrak{L}$ is locally finite and a refinement of $\mathfrak{S}$.

Proof: It is clear $\mathfrak{L}$ is a refinement of $\mathfrak{S}$ because every set of $\mathfrak{L}$ is a subset of a set of $\mathfrak{S}$, $F(B)$. Let $p \in S$. There exists an open set W, such that $p \in W$ and W intersects only $C_1, \cdots, C_n$, elements of $\mathfrak{C}$. Hence $W \subseteq \cup_{i=1}^n C_i$ since $\mathfrak{C}$ covers S.

But C_i is contained in a set $U_i \in \mathfrak{F}$ which intersects only finitely many sets of $\mathfrak{B}$. Thus each C_i intersects only finitely many $B \in \mathfrak{B}$ and so each C_i intersects only finitely many of the sets, $E(B)$. (If it intersects $E(B)$, then it intersects B.) Thus W intersects only finitely many of the $E(B)$, hence finitely many of the $E(B) \cap F(B)$. It follows that $\mathfrak{L}$ is locally finite.

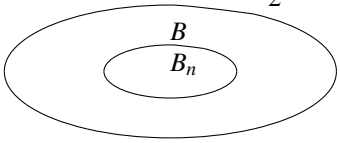
It is obvious that 4.)$\Rightarrow$ 1.). ■

The following theorem is Stone's theorem.

Theorem 25.0.6 *If S is a metric space then S is paracompact (Every open cover has a locally finite open refinement also an open cover.)*

Proof: Let $\mathfrak{S}$ be an open cover. Well order $\mathfrak{S}$. For $B \in \mathfrak{S}$,

$$B_n \equiv \{x \in B : \text{dist}\left(x, B^C\right) < \frac{1}{2^n}\},\ n = 1, 2, \cdots.$$



Thus B_n is contained in B but approximates it up to 2^{-n}. Let

$$E_n(B) = B_n \setminus \cup\{D : D \prec B \text{ and } D \neq B\}$$

where $\prec$ denotes the well order. If $B,\ D \in \mathfrak{S}$, then one is first in the well order. Let $D \prec B$. Then from the construction, $E_n(B) \subseteq D^C$ and $E_n(D)$ is further than $1/2^n$ from D^C. Hence, assuming neither set is empty,

$$\operatorname{dist}(E_n(B), E_n(D)) \geq 2^{-n}$$

for all $B,\ D \in \mathfrak{S}$. Fatten up $E_n(B)$ as follows.

$$\widetilde{E_n(B)} \equiv \cup\{B\left(x, 8^{-n}\right) : x \in E_n(B)\}.$$

Thus $\widetilde{E_n(B)} \subseteq B$ and

$$\operatorname{dist}\left(\widetilde{E_n(B)}, \widetilde{E_n(D)}\right) \geq \frac{1}{2^n} - 2\left(\frac{1}{8}\right)^n \equiv \delta_n > 0.$$

It follows that the collection of open sets

$$\{\widetilde{E_n(B)} : B \in \mathfrak{S}\} \equiv \mathfrak{B}_n$$

is locally finite. In fact, $B\left(p, \frac{\delta_n}{2}\right)$ cannot intersect more than one of them. In addition to this,

$$S \subseteq \cup\{\widetilde{E_n(B)} : n \in \mathbb{N},\ B \in \mathfrak{S}\}$$

because if $p \in S$, let B be the first set in $\mathfrak{S}$ to contain p. Then $p \in E_n(B)$ for n large enough because it will not be in anything deleted. Thus this is an open countably locally finite refinement. Thus 1.) in the above theorem is satisfied. ■

25.1 Partitions of Unity and Stone's Theorem

First recall that if S is nonempty, then $\operatorname{dist}(x,S)$ satisfies $|\operatorname{dist}(x,S) - \operatorname{dist}(y,S)| \leq d(x,y)$. It was Lemma 3.12.1.

Theorem 25.1.1 *Let S be a metric space and let $\mathfrak{S}$ be any open cover of S. Then there exists a set $\mathfrak{F}$, an open refinement of $\mathfrak{S}$, and functions $\{\phi_F : F \in \mathfrak{F}\}$ such that*

$$\phi_F : S \to [0,1]$$

$$\phi_F \text{ is continuous}$$

$$\phi_F(x) \text{ equals } 0 \text{ for all but finitely many } F \in \mathfrak{F}$$

$$\sum\{\phi_F(x) : F \in \mathfrak{F}\} = 1 \text{ for all } x \in S.$$

Each ϕ_F is locally Lipschitz continuous which means that for each z there is an open set W containing z for which, if $x, y \in W$, then there is a constant K such that

$$|\phi_F(x) - \phi_F(y)| \leq K d(x,y)$$

Proof: By Stone's theorem, there exists a locally finite open refinement $\mathfrak{F}$ of $\mathfrak{S}$ covering S. For $F \in \mathfrak{F}$

$$g_F(x) \equiv \operatorname{dist}\left(x, F^C\right)$$

Let

$$\phi_F(x) \equiv \left(\sum\{g_F(x) : F \in \mathfrak{F}\}\right)^{-1} g_F(x).$$

Now

$$\sum\{g_F(x) : F \in \mathfrak{F}\}$$

is a continuous function because if $x \in S$, then there exists an open set W with $x \in W$ and W has nonempty intersection with only finitely many sets of $F \in \mathfrak{F}$. Then for $y \in W$,

$$\sum\{g_F(y) : F \in \mathfrak{F}\} = \sum_{i=1}^{n} g_{F_i}(y).$$

Since $\mathfrak{F}$ is a cover of S,

$$\sum\{g_F(x) : F \in \mathfrak{F}\} \neq 0$$

for any $x \in S$. Hence ϕ_F is continuous. This also shows $\phi_F(x) = 0$ for all but finitely many $F \in \mathfrak{F}$. It is obvious that

$$\sum\{\phi_F(x) : F \in \mathfrak{F}\} = 1$$

from the definition.

Let $z \in S$. Then there is an open set W containing z such that W has nonempty intersection with only finitely many $F \in \mathscr{F}$. Thus for $y, x \in W$,

$$\left|\phi_{F_j}(x) - \phi_{F_j}(y)\right| \leq \left|\frac{g_{F_j}(x)\sum_{i=1}^{n} g_{F_i}(y) - g_{F_j}(y)\sum_{i=1}^{n} g_{F_i}(x)}{\sum_{i=1}^{n} g_{F_i}(x)\sum_{i=1}^{n} g_{F_i}(y)}\right|$$

If F is not one of these F_i, then $g_F(x) = \phi_F(x) = \phi_F(y) = g_F(y) = 0$. Thus there is nothing to show for these. It suffices to consider the ones above. Restricting W if necessary, we can assume that for $x \in W$,

$$\sum_F g_F(x) = \sum_{i=1}^{n} g_{F_i}(x) > \delta > 0,\ g_{F_j}(x) < \Delta < \infty,\ j \leq n$$

Then, simplifying the above, and letting $x, y \in W$, for each $j \leq n$,

$$\left|\phi_{F_j}(x) - \phi_{F_j}(y)\right| \leq \frac{1}{\delta^2}\left|\begin{array}{c} g_{F_j}(x)\sum_F g_F(y) - g_{F_j}(y)\sum_F g_F(y) \\ +g_{F_j}(y)\sum_F g_F(y) - g_{F_j}(y)\sum_F g_F(x) \end{array}\right|$$

$$\begin{aligned} &\leq \frac{1}{\delta^2}\Delta\left|g_{F_j}(x) - g_{F_j}(y)\right| + \frac{1}{\delta^2}\Delta\sum_{i=1}^{n}\left|g_{F_i}(y) - g_{F_i}(x)\right| \\ &\leq \frac{\Delta}{\delta^2}d(x,y) + \frac{\Delta}{\delta^2}nd(x,y) = (n+1)\frac{\Delta}{\delta^2}d(x,y) \end{aligned}$$

Thus on this set W containing z, all ϕ_F are Lipschitz continuous with Lipschitz constant $(n+1)\frac{\Delta}{\delta^2}$. ■

The functions described above are called a partition of unity subordinate to the open cover $\mathfrak{S}$. A useful observation is contained in the following corollary.

Corollary 25.1.2 *Let S be a metric space and let $\mathcal{S}$ be any open cover of S. Then there exists a set $\mathcal{F}$, an open refinement of $\mathcal{S}$, and functions $\{\phi_F : F \in \mathcal{F}\}$ such that*

$$\phi_F : S \to [0,1]$$

$$\phi_F \text{ is continuous}$$

$$\phi_F(x) \text{ equals } 0 \text{ for all but finitely many } F \in \mathcal{F}$$

$$\sum\{\phi_F(x) : F \in \mathcal{F}\} = 1 \text{ for all } x \in S.$$

Each ϕ_F is Lipschitz continuous. If $U \in \mathcal{S}$ and H is a closed subset of U, the partition of unity can be chosen such that each $\phi_F = 0$ on H except for one which equals 1 on H.

Proof: Just change your open cover to consist of U and $V \setminus H$ for each $V \in \mathcal{S}$. Then every function but one equals 0 on H and so exactly one of them equals 1 on H. ■

25.2 An Extension Theorem, Retracts

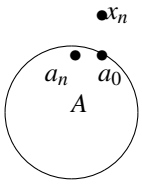
Recall the Tietze extension theorem which involved extending a real valued function defined on a closed set to a real valued function defined on the whole space. There is a big generalization in which the continuous function has values in a normed linear space. As with the Tietze extension theorem, the closed set is in a metric space.

Lemma 25.2.1 *Let A be a closed set in a metric space and let $x_n \notin A, x_n \to a_0 \in A$ and $a_n \in A$ such that $d(a_n, x_n) \leq 6\,\mathrm{dist}(x_n, A)$. Then $a_n \to a_0$.*

Proof: By assumption,

$$\begin{aligned} d(a_n, a_0) &\leq d(a_n, x_n) + d(x_n, a_0) < 6\,\mathrm{dist}(x_n, A) + d(x_n, a_0) \\ &\leq 6d(x_n, a_0) + d(x_n, a_0) = 7d(x_n, a_0) \end{aligned}$$

and this converges to 0. ■



In the proof of the following theorem, you get a covering of A^C with open balls B such that for each of these balls, there exists $a \in A$ such that for all $x \in B$, $\|x - a\|$ is no more than six times the distance of x to A. The 6 is not important. Any other constant with this property would work. Then you use Stone's theorem.

Recall a Banach space is a normed vector space which is also a complete metric space where the metric comes from the norm.

$$d(x,y) = \|x - y\|$$

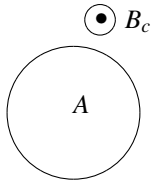
Thus you can add things in a Banach space.

Definition 25.2.2 *A Banach space is a complete normed linear space. If you have a subset B of a Banach space, then* $\operatorname{conv}(B)$ *denotes the smallest closed convex set which contains B. It can be obtained by taking the intersection of all closed convex sets containing B. Recall that a set C is convex if whenever $x,y \in C$, then so is $\lambda x + (1-\lambda) y$ for all $\lambda \in [0,1]$. Note how this makes sense in a vector space but maybe not in a general metric space.*

In the following theorem, we have in mind both X and Y are Banach spaces, but this is not needed in the proof. All that is needed is that X is a metric space and Y a normed linear space or possibly something more general in which it makes sense to do addition and scalar multiplication.

Theorem 25.2.3 *Let A be a closed subset of a metric space X and let $F : A \to Y$, Y a normed linear space. Then there exists an extension of F denoted as $\hat{F}$ such that $\hat{F}$ is defined on all of X and agrees with F on A. It has values in* $\operatorname{conv}(F(A))$, *the convex hull of $F(A)$.*

Proof: For each $c \notin A$, let B_c be a ball contained in A^C centered at c where distance of c to A is at least $\operatorname{diam}(B_c)$.



So for $x \in B_c$ what about $\operatorname{dist}(x,A)$? How does it compare with $\operatorname{dist}(c,A)$?

$$\begin{aligned} \operatorname{dist}(c,A) &\leq d(c,x) + \operatorname{dist}(x,A) \leq \frac{1}{2}\operatorname{diam}(B_c) + \operatorname{dist}(x,A) \\ &\leq \frac{1}{2}\operatorname{dist}(c,A) + \operatorname{dist}(x,A) \end{aligned}$$

so $\operatorname{dist}(c,A) \leq 2\operatorname{dist}(x,A)$. Now the following is also valid. Letting $x \in B_c$ be arbitrary, it follows from the assumption on the diameter that there exists $a_0 \in A$ such that $d(c,a_0) < 2\operatorname{dist}(c,A)$. Then

$$d(x,a_0) \leq \sup_{y \in B_c} d(y,a_0) \leq \sup_{y \in B_c} (d(y,c) + d(c,a_0)) \leq \frac{\operatorname{diam}(B_c)}{2} + 2\operatorname{dist}(c,A)$$

$$\leq \frac{\operatorname{dist}(c,A)}{2} + 2\operatorname{dist}(c,A) < 3\operatorname{dist}(c,A) \tag{25.1}$$

It follows from 25.1, $d(x,a_0) \leq 3\operatorname{dist}(c,A) \leq 6\operatorname{dist}(x,A)$. Thus for any $x \in B_c$, there is an $a_0 \in A$ such that $d(x,a_0)$ is bounded by a fixed multiple of the distance from x to A.

By Stone's theorem, there is a locally finite open refinement $\mathscr{R}$. These are open sets each of which is contained in one of the balls just mentioned such that each of these balls is the union of sets of $\mathscr{R}$. Thus $\mathscr{R}$ is a locally finite cover of A^C. Since $x \in A^C$ is in one of those balls, it was just shown that there exists $a_R \in A$ such that for all $x \in R \in \mathscr{R}$ we have $d(x,a_R) \leq 6\operatorname{dist}(x,A)$. Of course there may be more than one because R might be contained in more than one of those special balls. One a_R is chosen for each $R \in \mathscr{R}$.

Now let $\phi_R(x) \equiv \text{dist}\left(x, R^C\right)$. Then let

$$\hat{F}(x) \equiv \begin{cases} F(x) \text{ for } x \in A \\ \sum_{R\in\mathscr{R}} F(a_R) \frac{\phi_R(x)}{\sum_{\hat{R}\in\mathscr{R}} \phi_{\hat{R}}(x)} \text{ for } x \notin A \end{cases}$$

The sum in the bottom is always finite because the covering is locally finite. Also, this sum is never 0 because $\mathscr{R}$ is a covering. Also $\hat{F}$ has values in $\text{conv}(F(K))$. It only remains to verify that $\hat{F}$ is continuous. It is clearly so on the interior of A thanks to continuity of F. It is also clearly continuous on A^C because the functions ϕ_R are continuous. So it suffices to consider $x_n \to a \in \partial A \subseteq A$ where $x_n \notin A$ and see whether $F(a) = \lim_{n\to\infty} \hat{F}(x_n)$.

Suppose this does not happen. Then there is a sequence converging to some $a \in \partial A$ and $\varepsilon > 0$ such that

$$\varepsilon \leq \left\|\hat{F}(a) - \hat{F}(x_n)\right\| \text{ all } n$$

For $x_n \in R$, it was shown above that $d(x_n, a_{R_n}) \leq 6\,\text{dist}(x_n, A)$. By the above Lemma 25.2.1, it follows that $a_{Rn} \to a$ and so $F(a_{Rn}) \to F(a)$.

$$\varepsilon \leq \left\|\hat{F}(a) - \hat{F}(x_n)\right\| \leq \sum_{R\in\mathscr{R}} \|F(a_{Rn}) - F(a)\| \frac{\phi_R(x_{Rn})}{\sum_{\hat{R}\in\mathscr{R}} \phi_{\hat{R}}(x_{Rn})}$$

By local finiteness of the cover, each x_n involves only finitely many R Thus, in this limit process, there are countably many R involved $\left\{R_j\right\}_{j=1}^{\infty}$. Thus one can apply Fatou's lemma.

$$\begin{aligned} \varepsilon &\leq \liminf_{n\to\infty} \left\|\hat{F}(a) - \hat{F}(x_n)\right\| \\ &\leq \sum_{j=1}^{\infty} \liminf_{n\to\infty} \left\|F\left(a_{R_jn}\right) - F(a)\right\| \frac{\phi_{R_j}\left(x_{R_jn}\right)}{\sum_{j=1}^{\infty} \phi_{\hat{R}_j}\left(x_{R_jn}\right)} \\ &\leq \sum_{j=1}^{\infty} \liminf_{n\to\infty} \left\|F\left(a_{R_jn}\right) - F(a)\right\| = 0 \blacksquare \end{aligned}$$

The last step is needed because you lose local finiteness as you approach ∂A. Note that the only thing needed was that X is a metric space. The addition takes place in Y so it needs to be a vector space. Did it need to be complete? No, this was not used. Nor was completeness of X used. The main interest here is in Banach spaces, but the result is more general than that.

It also appears that $\hat{F}$ is locally Lipschitz on A^C.

Definition 25.2.4 *Let S be a subset of X, a Banach space. Then it is a retract if there exists a continuous function $R : X \to S$ such that $Rs = s$ for all $s \in S$. This R is a retraction. More generally, $S \subseteq T$ is called a retract of T if there is a continuous $R : T \to S$ such that $Rs = s$ for all $s \in S$.*

Theorem 25.2.5 *Let K be closed and convex subset of X a Banach space. Then K is a retract.*

Proof: By Theorem 25.2.3, there is a continuous function $\hat{I}$ extending I to all of X. Then also $\hat{I}$ has values in $\text{conv}(IK) = \text{conv}(K) = K$. Hence $\hat{I}$ is a continuous function which does what is needed. It maps everything into K and keeps the points of K unchanged. $\blacksquare$

Sometimes people call the set a retraction also or the function which does the job a retraction. This seems like strange thing to call it because a retraction is the act of repudiating something you said earlier. Nevertheless, I will call it that. Note that if S is a retract of the whole metric space X, then it must be a retract of every set which contains S.

Part IV

Stochastic Processes and Probability

Chapter 26

Independence

Caution: This material on probability and stochastic processes may be half baked in places. This is not to say that nothing else is half baked. However, the probability is higher here. Probability is not my main research area so my qualifications for even writing this are not all they could be. However, I like probability and think it fits in well with what is presented earlier and hope this might be useful for someone like me. This book is not a research monograph written for experts, and I am no expert. I have included mainly those items which I have found most interesting. However, there is an awful lot in this subject, far more than I can include. For more topics see the references.

This material was written down earlier in the Topics in Analysis book but in a haphazard manner as I encountered it rather than in the most logical manner. I am trying to present it here in a more coherent form.

26.1 Random Variables and Independence

Recall Lemma 20.2.3 on Page 548 which is stated here for convenience.

Lemma 26.1.1 *Let M be a metric space with the closed balls compact and suppose λ is a measure defined on the Borel sets of M which is finite on compact sets. Then there exists a unique Radon measure, $\overline{\lambda}$ which equals λ on the Borel sets. In particular λ must be both inner and outer regular on all Borel sets.*

Also important is the following fundamental result which is called the Borel Cantelli lemma. It is Lemma 9.2.5 on Page 247.

Lemma 26.1.2 *Let $(\Omega, \mathscr{F}, \lambda)$ be a measure space and let $\{A_i\}$ be a sequence of measurable sets satisfying $\sum_{i=1}^{\infty} \lambda(A_i) < \infty$. Then letting S denote the set of $\omega \in \Omega$ which are in infinitely many A_i, it follows S is a measurable set and $\lambda(S) = 0$.*

Here is another nice observation.

Proposition 26.1.3 *Suppose E_i is a separable Banach space. Then if B_i is a Borel set of E_i, it follows $\prod_{i=1}^n B_i$ is a Borel set in $\prod_{i=1}^n E_i$.*

Proof: An easy way to do this is to consider the projection maps $\pi_i \boldsymbol{x} \equiv x_i$. Then these projection maps are continuous. Hence for U open, $\pi_i^{-1}(U) \equiv \prod_{j=1}^n A_j$, $A_j = E_j$ if $j \neq i$ and $A_i = U$. Thus $\pi_i^{-1}(open)$ equals an open set. Let $\mathscr{S} \equiv \left\{V \subseteq \mathbb{R} : \pi_i^{-1}(V) \text{ is Borel}\right\}$. Then $\mathscr{S}$ contains all the open sets and is clearly a σ algebra. Therefore, $\mathscr{S}$ contains the Borel sets. Let B_i be a Borel set in E_i. Then $\prod_{i=1}^n B_i = \cap_{i=1}^n \pi_i^{-1}(B_i)$, a finite intersection of Borel sets. ■

Definition 26.1.4 *A probability space is a measure space, $(\Omega, \mathscr{F}, P)$ where P is a measure satisfying $P(\Omega) = 1$. A random vector (variable) is a measurable function, $\boldsymbol{X} : \Omega \to Z$ where Z is some topological space. It is often the case that Z will equal $\mathbb{R}^p$. Assume Z is a separable Banach space. Define the following σ algebra.*

$$\sigma(\boldsymbol{X}) \equiv \left\{\boldsymbol{X}^{-1}(E) : E \text{ is Borel in } Z\right\}$$

Thus $\sigma(\boldsymbol{X}) \subseteq \mathscr{F}$. For E a Borel set in Z define $\lambda_{\boldsymbol{X}}(E) \equiv P\left(\boldsymbol{X}^{-1}(E)\right)$. This is called the distribution of the random variable $\boldsymbol{X}$. If $\int_\Omega |\boldsymbol{X}(\omega)| dP < \infty$ then define $E(\boldsymbol{X}) \equiv \int_\Omega \boldsymbol{X} dP$ where the integral is defined as the Bochner integral.

Recall the following fundamental result which was proved earlier but which I will give a short proof of now.

Proposition 26.1.5 *Let $(\Omega,\mathscr{S},\mu)$ be a measure space and let $\boldsymbol{X}:\Omega\to Z$ where Z is a separable Banach space. Then $\boldsymbol{X}$ is strongly measurable if and only if $\boldsymbol{X}^{-1}(U)\in\mathscr{S}$ for all U open in Z.*

Proof: To begin with, let $D(a,r)$ be the closure of the open ball $B(a,r)$. By Lemma 24.1.7, there exists $\{f_i\}\subseteq B'$, the unit ball in Z' such that $\|z\|_Z=\sup_i\{|f_i(z)|\}$. Then

$$D(a,r)=\{z:\|a-z\|\le r\}=\cap_i\{z:|f_i(z)-f_i(a)|\le r\}$$

$$=\cap_i f_i^{-1}\left(\overline{B(f_i(a),r)}\right)$$

Thus $\boldsymbol{X}^{-1}(D(a,r))=\cap_i\boldsymbol{X}^{-1}\left(f_i^{-1}\left(\overline{B(f_i(a),r)}\right)\right)=\cap_i(f_i\circ\boldsymbol{X})^{-1}\left(\overline{B(f_i(a),r)}\right)$. If $\boldsymbol{X}$ is strongly measurable, then it is weakly measurable and so each $f_i\circ\boldsymbol{X}$ is a real (complex) valued measurable function. Hence the expression on the right in the above is measurable. Now if U is any open set in Z, then it is the countable union of such closed disks $U=\cup_i D_i$. Therefore, $\boldsymbol{X}^{-1}(U)=\cap_i\boldsymbol{X}^{-1}(D_i)\in\mathscr{S}$. It follows that strongly measurable implies inverse images of open sets are in $\mathscr{S}$.

Conversely, suppose $\boldsymbol{X}^{-1}(U)\in\mathscr{S}$ for every open U. Then for $f\in Z'$, $f\circ\boldsymbol{X}$ is real valued and measurable. Therefore, $\boldsymbol{X}$ is weakly measurable. By the Pettis theorem, it follows that $f\circ\boldsymbol{X}$ is strongly measurable. ■

Proposition 26.1.6 *If $\boldsymbol{X}:\Omega\to Z$ is measurable, then $\sigma(\boldsymbol{X})$ equals the smallest σ algebra such that $\boldsymbol{X}$ is measurable with respect to it. Also if X_i are random variables having values in separable Banach spaces Z_i, then $\sigma(\boldsymbol{X})=\sigma(X_1,\cdots,X_n)$ where $\boldsymbol{X}$ is the vector mapping Ω to $\prod_{i=1}^n Z_i$ and $\sigma(X_1,\cdots,X_n)$ is the smallest σ algebra such that each X_i is measurable with respect to it.*

Proof: Let $\mathscr{G}$ denote the smallest σ algebra such that $\boldsymbol{X}$ is measurable with respect to this σ algebra. By definition $\boldsymbol{X}^{-1}(\text{open})\in\mathscr{G}$. Furthermore, the set of all E such that $\boldsymbol{X}^{-1}(E)\in\mathscr{G}$ is a σ algebra. Hence it includes all the Borel sets. Hence $\boldsymbol{X}^{-1}(\text{Borel})\in\mathscr{G}$ and so $\mathscr{G}\supseteq\sigma(\boldsymbol{X})$. However, $\sigma(\boldsymbol{X})$ defined above is a σ algebra such that $\boldsymbol{X}$ is measurable with respect to $\sigma(\boldsymbol{X})$. Therefore, $\mathscr{G}=\sigma(\boldsymbol{X})$.

Letting B_i be a Borel set in Z_i, $\prod_{i=1}^n B_i$ is a Borel set by Proposition 26.1.3 and so $\boldsymbol{X}^{-1}(\prod_{i=1}^n B_i)=\cap_{i=1}^n X_i^{-1}(B_i)\in\sigma(X_1,\cdots,X_n)$. If $\mathscr{G}$ denotes the Borel sets $F\subseteq\prod_{i=1}^n Z_i$ such that $\boldsymbol{X}^{-1}(F)\in\sigma(X_1,\cdots,X_n)$, then $\mathscr{G}$ is clearly a σ algebra which contains the open sets. Hence $\mathscr{G}=\mathscr{B}$ the Borel sets of $\prod_{i=1}^n Z_i$. This shows that $\sigma(\boldsymbol{X})\subseteq\sigma(X_1,\cdots,X_n)$. Next we observe that $\sigma(\boldsymbol{X})$ is a σ algebra with the property that each X_i is measurable with respect to $\sigma(\boldsymbol{X})$. This follows from $X_i^{-1}(B_i)=\boldsymbol{X}^{-1}\left(\prod_{j=1}^n A_j\right)\in\sigma(\boldsymbol{X})$, where each $A_j=Z_j$ except for $A_i=B_i$.Since $\sigma(X_1,\cdots,X_n)$ is defined as the smallest such σ algebra, it follows that $\sigma(\boldsymbol{X})\supseteq\sigma(X_1,\cdots,X_n)$.■

For random variables having values in a separable Banach space or even more generally for a separable metric space, much can be said about regularity of $\lambda_{\boldsymbol{X}}$.

Definition 26.1.7 *A measure, μ defined on $\mathscr{B}(E)$ for E a separable metric space will be called inner regular if for all $F\in\mathscr{B}(E)$,*

$$\mu(F)=\sup\{\mu(K):K\subseteq F\text{ and }K\text{ is closed}\}$$

A measure, μ defined on $\mathscr{B}(E)$ will be called outer regular if for all $F \in \mathscr{B}(E)$,

$$\mu(F) = \inf\{\mu(V) : V \supseteq F \text{ and } V \text{ is open}\}$$

When a measure is both inner and outer regular, it is called regular.

Note that if the metric space is $\mathbb{R}^p$ then $\lambda_{\boldsymbol{X}}$ can be considered a Radon measure because you can use it to obtain a positive linear functional and then use the Riesz representation theorem for these.

For probability measures, the above definition of regularity tends to come free. Note it is a little weaker than the usual definition of regularity because K is only assumed to be closed, not compact. This is stated for convenience. It is Lemma 9.4.4 on Page 250.

Lemma 26.1.8 *Let μ be a finite measure defined on $\mathscr{B}(E)$ where E is a metric space. Then μ is regular.*

One can say more if the metric space is complete and separable. In fact in this case the above definition of inner regularity can be shown to imply the usual one where the closed sets are replaced with compact sets. It is Lemma 9.4.5 on Page 251.

Lemma 26.1.9 *Let μ be a finite measure on a σ algebra containing $\mathscr{B}(X)$, the Borel sets of X, a separable complete metric space. (Polish space) Then if C is a closed set,*

$$\mu(C) = \sup\{\mu(K) : K \subseteq C \text{ and } K \text{ is compact.}\}$$

It follows that for a finite measure on $\mathscr{B}(X)$ where X is a Polish space, μ is inner regular in the sense that for all $F \in \mathscr{B}(X)$,

$$\mu(F) = \sup\{\mu(K) : K \subseteq F \text{ and } K \text{ is compact}\}$$

Definition 26.1.10 *A measurable function $\boldsymbol{X} : (\Omega, \mathscr{F}, \mu) \to Z$ a topological space is called a random variable when $\mu(\Omega) = 1$. For such a random variable, one can define a distribution measure $\lambda_{\boldsymbol{X}}$ on the Borel sets of Z as follows: $\lambda_{\boldsymbol{X}}(G) \equiv \mu\left(\boldsymbol{X}^{-1}(G)\right)$. This is a well defined measure on the Borel sets of Z because it makes sense for every G open and $\mathscr{G} \equiv \left\{G \subseteq Z : \boldsymbol{X}^{-1}(G) \in \mathscr{F}\right\}$ is a σ algebra which contains the open sets, hence the Borel sets. Such a measurable function is also called a random vector.*

Corollary 26.1.11 *Let $\boldsymbol{X}$ be a random variable (random vector) with values in a complete metric space, Z. Then $\lambda_{\boldsymbol{X}}$ is an inner and outer regular measure defined on $\mathscr{B}(Z)$.*

Proposition 26.1.12 *For $\boldsymbol{X}$ a random vector defined above, $\boldsymbol{X}$ having values in a complete separable metric space Z, then $\lambda_{\boldsymbol{X}}$ is inner and outer regular and Borel.*

$$(\Omega, P) \xrightarrow{\boldsymbol{X}} (Z, \lambda_{\boldsymbol{X}}) \xrightarrow{h} E$$

If h is Borel measurable and $h \in L^1(Z, \lambda_{\boldsymbol{X}}; E)$ for E a Banach space, then

$$\int_\Omega h(\boldsymbol{X}(\omega))\, dP = \int_Z h(\boldsymbol{x})\, d\lambda_{\boldsymbol{X}}. \tag{26.1}$$

In the case where $Z = E$, a separable Banach space, if $\boldsymbol{X}$ is measurable then $\boldsymbol{X} \in L^1(\Omega; E)$ if and only if the identity map on E is in $L^1(E; \lambda_{\boldsymbol{X}})$ and

$$\int_\Omega \boldsymbol{X}(\omega)\, dP = \int_E \boldsymbol{x} d\lambda_{\boldsymbol{X}}(\boldsymbol{x}) \tag{26.2}$$

Proof: The regularity claims are established above. It remains to verify 26.1.

Since $h \in L^1(Z,E)$, it follows there exists a sequence of simple functions $\{h_n\}$ such that

$$h_n(\boldsymbol{x}) \to h(\boldsymbol{x}), \int_Z \|h_m - h_n\| \, d\lambda_{\boldsymbol{X}} \to 0 \text{ as } m,n \to \infty.$$

The first convergence above implies

$$h_n \circ \boldsymbol{X} \to h \circ \boldsymbol{X} \text{ pointwise on } \Omega \tag{26.3}$$

Then letting $h_n(\boldsymbol{x}) = \sum_{k=1}^m \boldsymbol{x}_k \mathcal{X}_{E_k}(\boldsymbol{x})$, where the E_k are disjoint and Borel, it follows easily that $h_n \circ \boldsymbol{X}$ is also a simple function of the form $h_n \circ \boldsymbol{X}(\omega) = \sum_{k=1}^m \boldsymbol{x}_k \mathcal{X}_{\boldsymbol{X}^{-1}(E_k)}(\omega)$ and by assumption $\boldsymbol{X}^{-1}(E_k) \in \mathcal{F}$. From the definition of the integral, it is easily seen

$$\int h_n \circ \boldsymbol{X} \, dP = \int h_n \, d\lambda_{\boldsymbol{X}}, \int \|h_n\| \circ \boldsymbol{X} \, dP = \int \|h_n\| \, d\lambda_{\boldsymbol{X}}$$

Also, $h_n \circ \boldsymbol{X} - h_m \circ \boldsymbol{X}$ is a simple function and so

$$\int \|h_n \circ \boldsymbol{X} - h_m \circ \boldsymbol{X}\| \, dP = \int \|h_n - h_m\| \, d\lambda_{\boldsymbol{X}} \tag{26.4}$$

It follows from the definition of the Bochner integral and 26.3, and 26.4 that $h \circ \boldsymbol{X}$ is in $L^1(\Omega;E)$ and

$$\int h \circ \boldsymbol{X} \, dP = \lim_{n\to\infty} \int h_n \circ \boldsymbol{X} \, dP = \lim_{n\to\infty} \int h_n \, d\lambda_{\boldsymbol{X}} = \int h \, d\lambda_{\boldsymbol{X}}.$$

Finally consider the case that $E = Z$ for E a separable Banach space, and suppose $\boldsymbol{X} \in L^1(\Omega;E)$. Then letting h be the identity map on E, it follows h is obviously separably valued and $h^{-1}(U) \in \mathcal{B}(E)$ for all U open and so h is measurable. Why is it in $L^1(E;E)$?

$$\begin{aligned} \int_E \|h(\boldsymbol{x})\| \, d\lambda_{\boldsymbol{X}} &= \int_0^\infty \lambda_{\boldsymbol{X}}([\|h\| > t]) \, dt \equiv \int_0^\infty P(\boldsymbol{X} \in [\|\boldsymbol{x}\| > t]) \, dt \\ &\equiv \int_0^\infty P([\|\boldsymbol{X}\| > t]) \, dt = \int_\Omega \|\boldsymbol{X}\| \, dP < \infty \end{aligned}$$

Thus the identity map on E is in $L^1(E;\lambda_{\boldsymbol{X}})$. Next let the identity map h be in $L^1(E;\lambda_{\boldsymbol{X}})$. Then $\boldsymbol{X}(\omega) = h \circ \boldsymbol{X}(\omega)$ and so from the first part, $\boldsymbol{X} \in L^1(\Omega;E)$ and from 26.1, 26.2 follows. ■

26.2 Convergence in Probability

Definition 26.2.1 *$\{f_n\}$ is said to be Cauchy in probability if for each $\varepsilon > 0$,*

$$\lim_{n,m\to\infty} P(\|f_n - f_m\| > \varepsilon) = 0$$

This means: for each $\delta > 0$ there exists k_δ such that if $m,n \geq k_\delta$, then $P(\|f_n - f_m\| > \varepsilon) < \delta$.

Proposition 26.2.2 *$\{f_n\}$ is Cauchy in probability, these being functions having values in a Banach space then there exists a set of measure zero N and a subsequence $\{f_{n_k}\}$ such that for $\omega \notin N$, $\lim_{k\to\infty} f_{n_k}(\omega)$ converges.*

Proof: From the above definition, there exists n_1 such that if $m,n \geq n_1$, then

$$P\left(\|f_n - f_m\| > 2^{-1}\right) < 2^{-1}$$

From the definition, there exists $n_2 > n_1$ such that $P\left(\|f_n - f_m\| > 2^{-2}\right) < 2^{-2}$ whenever $n,m \geq n_1$ and so forth. Thus

$$P\left(\left\|f_{n_{k+1}} - f_{n_k}\right\| > 2^{-k}\right) < 2^{-k}$$

Letting $A_k \equiv \left[\left\|f_{n_{k+1}} - f_{n_k}\right\| > 2^{-k}\right]$, it follows from the Borell Cantelli lemma that there is a set of measure zero, namely $N \equiv \cap_{n=1}^{\infty} \cup_{k \geq n} A_k$ such that $P\left(N^C\right) = 1 = P\left(\cup_{n=1}^{\infty} \cap_{k \geq n} A_k^C\right)$. To say $\omega \in N^C$ is the same as saying that there exists n such that ω is in A_k^C for all $k \geq n$. In other words, eventually $\left\|f_{n_{k+1}} - f_{n_k}\right\| \leq 2^{-k}$. Now it follows that

$$\left\|f_{n+p}(\omega) - f_{n_k}(\omega)\right\| \leq \sum_{j=k}^{\infty} \left\|f_{n+j+1}(\omega) - f_{n+j}(\omega)\right\| < 2^{-(k-1)}$$

if k is large enough. Hence $\left\{f_{n_k}(\omega)\right\}_k$ is a Cauchy sequence for each $\omega \notin N$ and since E is complete, this sequence converges. ■

26.3 Kolmogorov Extension Theorem

Let M_t be a complete separable metric space. This is called a Polish space. I will denote a totally ordered index set, (Like $\mathbb{R}$) and the interest will be in building a measure on the product space, $\prod_{t \in I} M_t$. If you like less generality, just think of $M_t = \mathbb{R}^{k_t}$ or even $M_t = \mathbb{R}$. By the well ordering principle, you can always put an order on any index set so this order is no restriction, but we do not insist on a well order and in fact, index sets of great interest are $\mathbb{R}$ or $[0,\infty)$. Also for X a topological space, $\mathscr{B}(X)$ will denote the Borel sets.

Notation 26.3.1 *The symbol J will denote a finite subset of I, $J = (t_1,\cdots,t_n)$, the t_i taken in order. $\boldsymbol{E}_J$ will denote a set which has a set E_t of $\mathscr{B}(M_t)$ in the t^{th} position for $t \in J$ and for $t \notin J$, the set in the t^{th} position will be M_t. $\boldsymbol{K}_J$ will denote a set which has a compact set in the t^{th} position for $t \in J$ and for $t \notin J$, the set in the t^{th} position will be M_t. Also denote by $\mathscr{R}_J$ the sets $\boldsymbol{E}_J$ and $\mathscr{R}$ the union of all such $\mathscr{R}_J$. Let $\mathscr{E}_J$ denote finite disjoint unions of sets of $\mathscr{R}_J$ and let $\mathscr{E}$ denote finite disjoint unions of sets of $\mathscr{R}$. Thus if $\boldsymbol{F}$ is a set of $\mathscr{E}$, there exists J such that $\boldsymbol{F}$ is a finite disjoint union of sets of $\mathscr{R}_J$. For $\boldsymbol{F} \in \Omega$, denote by $\pi_J(\boldsymbol{F})$ the set $\prod_{t \in J} F_t$ where $\boldsymbol{F} = \prod_{t \in I} F_t$.*

With this preparation, here is the Kolmogorov extension theorem. It is Theorem 20.3.3 proved earlier. In the statement and proof of the theorem, F_i, G_i, and E_i will denote Borel sets. Any list of indices from I will always be assumed to be taken in order. Thus, if $J \subseteq I$ and $J = (t_1,\cdots,t_n)$, it will always be assumed $t_1 < t_2 < \cdots < t_n$.

Theorem 26.3.2 *For each finite set $J = (t_1,\cdots,t_n) \subseteq I$, suppose there exists a Borel probability measure, $\nu_J = \nu_{t_1\cdots t_n}$ defined on the Borel sets of $\prod_{t \in J} M_t$ such that the following consistency condition holds. If $(t_1,\cdots,t_n) \subseteq (s_1,\cdots,s_p)$, then*

$$\nu_{t_1\cdots t_n}(F_{t_1} \times \cdots \times F_{t_n}) = \nu_{s_1\cdots s_p}\left(G_{s_1} \times \cdots \times G_{s_p}\right) \tag{26.5}$$

where if $s_i = t_j$, then $G_{s_i} = F_{t_j}$ and if s_i is not equal to any of the indices t_k, then $G_{s_i} = M_{s_i}$. Then for $\mathscr{E}$ defined in the above Notation, there exists a probability measure P and a σ algebra $\mathscr{F} = \sigma(\mathscr{E})$ such that $(\prod_{t\in I} M_t, P, \mathscr{F})$ is a probability space. Also there exist measurable functions, $X_s : \prod_{t\in I} M_t \to M_s$ defined for $s \in I$ as

$$X_s \boldsymbol{x} \equiv x_s$$

such that for each $(t_1 \cdots t_n) \subseteq I$,

$$\begin{aligned} \nu_{t_1\cdots t_n}(F_{t_1} \times \cdots \times F_{t_n}) &= P([X_{t_1} \in F_{t_1}] \cap \cdots \cap [X_{t_n} \in F_{t_n}]) \\ &= P\left((X_{t_1}, \cdots, X_{t_n}) \in \prod_{j=1}^{n} F_{t_j}\right) = P\left(\prod_{t\in I} F_t\right) \end{aligned} \tag{26.6}$$

where $F_t = M_t$ for every $t \notin \{t_1 \cdots t_n\}$ and F_{t_i} is a Borel set. Also if f is a nonnegative function of finitely many variables, $x_{t_1}, \cdots, x_{t_n}$, measurable with respect to $\mathscr{B}\left(\prod_{j=1}^{n} M_{t_j}\right)$, then f is also measurable with respect to $\mathscr{F}$ and

$$\int_{M_{t_1}\times\cdots\times M_{t_n}} f(x_{t_1}, \cdots, x_{t_n})\, d\nu_{t_1\cdots t_n} = \int_{\prod_{t\in I} M_t} f(x_{t_1}, \cdots, x_{t_n})\, dP \tag{26.7}$$

26.4 Independent Events and σ Algebras

The concept of independence is probably the main idea which separates probability from analysis and causes some of us, myself included, to struggle to understand what is going on. I think that these ideas are the main difficulty some encounter when trying to understand probability, not the kind based on combinatorics but what is being presented here.

Definition 26.4.1 *Let $(\Omega, \mathscr{F}, P)$ be a probability space. The sets in $\mathscr{F}$ are called events. A set of events, $\{A_i\}_{i\in I}$ is called independent if whenever $\{A_{i_k}\}_{k=1}^{m}$ is a finite subset*

$$P\left(\cap_{k=1}^{m} A_{i_k}\right) = \prod_{k=1}^{m} P\left(A_{i_k}\right).$$

Each of these events defines a rather simple σ algebra, $\left(A_i, A_i^C, \emptyset, \Omega\right)$ denoted by $\mathscr{F}_i$. Now the following lemma is interesting because it motivates a more general notion of independent σ algebras.

Lemma 26.4.2 *Suppose $\{A_i\}_{i\in I}$ are independent events. Then for*

$$B_i \in \mathscr{F}_i \equiv \left\{A_i, A_i^C, \emptyset, \Omega\right\}$$

for $i \in I$. Then for any $m \in \mathbb{N}$, $P\left(\cap_{k=1}^{m} B_{i_k}\right) = \prod_{k=1}^{m} P\left(B_{i_k}\right)$.

Proof: The proof is by induction on the number l of the B_{i_k} which are not equal to A_{i_k}. First suppose $l = 0$. Then the above assertion is true by assumption since $\{A_i\}_{i\in I}$ is independent. Suppose it is so for some l and there are $l+1$ sets not equal to A_{i_k}. If any equals $\emptyset$ there is nothing to show. Both sides equal 0. If any equals Ω, there is also nothing to show. You can ignore that set in both sides and then you have by induction the two sides

are equal because you have no more than l sets different than A_{i_k}. The only remaining case is where some $B_{i_k} = A_{i_k}^C$. Say $B_{i_{m+1}} = A_{i_{m+1}}^C$ for simplicity.

$$P\left(\cap_{k=1}^{m+1} B_{i_k}\right) = P\left(A_{i_{m+1}}^C \cap \cap_{k=1}^{m} B_{i_k}\right) = P\left(\cap_{k=1}^{m} B_{i_k}\right) - P\left(A_{i_{m+1}} \cap \cap_{k=1}^{m} B_{i_k}\right)$$

Then by induction,

$$\begin{aligned} &= \prod_{k=1}^{m} P\left(B_{i_k}\right) - P\left(A_{i_{m+1}}\right) \prod_{k=1}^{m} P\left(B_{i_k}\right) = \prod_{k=1}^{m} P\left(B_{i_k}\right)\left(1 - P\left(A_{i_{m+1}}\right)\right) \\ &= P\left(A_{i_{m+1}}^C\right) \prod_{k=1}^{m} P\left(B_{i_k}\right) = \prod_{k=1}^{m+1} P\left(B_{i_k}\right) \end{aligned}$$

thus proving it for $l+1$. ■

This motivates a more general notion of independence in terms of σ algebras.

Definition 26.4.3 *If $\{\mathscr{F}_i\}_{i\in I}$ is any set of σ algebras contained in $\mathscr{F}$, they are said to be independent if whenever $A_{i_k} \in \mathscr{F}_{i_k}$ for $k = 1, 2, \cdots, m$, then*

$$P\left(\cap_{k=1}^{m} A_{i_k}\right) = \prod_{k=1}^{m} P\left(A_{i_k}\right).$$

A set of random variables $\{\boldsymbol{X}_i\}_{i\in I}$ is independent if the σ algebras $\{\sigma(\boldsymbol{X}_i)\}_{i\in I}$ are independent σ algebras. Here $\sigma(\boldsymbol{X})$ denotes the smallest σ algebra such that $\boldsymbol{X}$ is measurable. Thus $\sigma(\boldsymbol{X}) = \left\{\boldsymbol{X}^{-1}(U) : U \text{ is a Borel set}\right\}$. More generally, $\sigma(\boldsymbol{X}_i : i \in I)$ is the smallest σ algebra such that each $\boldsymbol{X}_i$ is measurable.

Note that by Lemma 26.4.2 you can consider independent events in terms of independent σ algebras. That is, a set of independent events can always be considered as events taken from a set of independent σ algebras. This is a more general notion because here the σ algebras might have infinitely many sets in them.

Lemma 26.4.4 *Suppose the set of random variables, $\{\boldsymbol{X}_i\}_{i\in I}$ is independent. Also suppose $I_1 \subseteq I$ and $j \notin I_1$. Then the σ algebras $\sigma(\boldsymbol{X}_i : i \in I_1)$, $\sigma(\boldsymbol{X}_j)$ are independent σ algebras.*

Proof: Let $B \in \sigma(\boldsymbol{X}_j)$. I want to show that for any $A \in \sigma(\boldsymbol{X}_i : i \in I_1)$, it follows that $P(A \cap B) = P(A)P(B)$. Let $\mathscr{K}$ consist of finite intersections of sets of the form $\boldsymbol{X}_k^{-1}(B_k)$ where B_k is a Borel set and $k \in I_1$. Thus $\mathscr{K}$ is a π system and $\sigma(\mathscr{K}) = \sigma(\boldsymbol{X}_i : i \in I_1)$. This is because it follows from the definition that $\sigma(\mathscr{K}) \supseteq \sigma(\boldsymbol{X}_i : i \in I_1)$ because $\sigma(\mathscr{K})$ contains all $\boldsymbol{X}_i^{-1}(B)$ for B Borel. For the other inclusion, the right side consists of all sets $\boldsymbol{X}_i^{-1}(B)$ where B is a Borel set and so the right side, being a σ algebra also must contain all finite intersections of these sets which means $\sigma(\boldsymbol{X}_i : i \in I_1)$ must contain $\mathscr{K}$ and so $\sigma(\boldsymbol{X}_i : i \in I_1) \supseteq \sigma(\mathscr{K})$.

Now if you have one of these sets of the form $A = \cap_{k=1}^{m} \boldsymbol{X}_k^{-1}(B_k)$ where without loss of generality, it can be assumed the k are distinct since $\boldsymbol{X}_k^{-1}(B_k) \cap \boldsymbol{X}_k^{-1}\left(B_k'\right) = \boldsymbol{X}_k^{-1}\left(B_k \cap B_k'\right)$, then

$$\begin{aligned} P(A \cap B) &= P\left(\cap_{k=1}^{m} \boldsymbol{X}_k^{-1}(B_k) \cap B\right) = P(B) \prod_{k=1}^{m} P\left(\boldsymbol{X}_k^{-1}(B_k)\right) \\ &= P(B)\, P\left(\cap_{k=1}^{m} \boldsymbol{X}_k^{-1}(B_k)\right). \end{aligned}$$

Thus $\mathscr{K}$ is contained in

$$\mathscr{G} \equiv \{A \in \sigma(\boldsymbol{X}_i : i \in I_1) : P(A \cap B) = P(A) P(B)\}.$$

Now $\mathscr{G}$ is closed with respect to complements and countable disjoint unions. Here is why: If each $A_i \in \mathscr{G}$ and the A_i are disjoint,

$$\begin{aligned} P((\cup_{i=1}^{\infty} A_i) \cap B) &= P(\cup_{i=1}^{\infty} (A_i \cap B)) \\ &= \sum_i P(A_i \cap B) = \sum_i P(A_i) P(B) \\ &= P(B) \sum_i P(A_i) = P(B) P(\cup_{i=1}^{\infty} A_i) \end{aligned}$$

If $A \in \mathscr{G}$, $P(A^C \cap B) + P(A \cap B) = P(B)$ and so

$$\begin{aligned} P(A^C \cap B) &= P(B) - P(A \cap B) = P(B) - P(A) P(B) \\ &= P(B)(1 - P(A)) = P(B) P(A^C). \end{aligned}$$

Therefore, from the lemma on π systems, Lemma 9.3.2 on Page 247, it follows

$$\sigma(\boldsymbol{X}_i : i \in I_1) \supseteq \mathscr{G} \supseteq \sigma(\mathscr{K}) = \sigma(\boldsymbol{X}_i : i \in I_1). \blacksquare$$

Definition 26.4.5 *When X is a random variable with values in a Banach space Z, the notation $E(X)$ means $\int_Z X(\omega) dP$ where the latter is the Bochner integral. I will use this notation whenever convenient. $E(X)$ is called the expectation of X or the expected value of X. I will sometimes also use E as the name of a Banach space, but it should be clear from the context which is meant.*

Lemma 26.4.6 *If $\{\boldsymbol{X}_k\}_{k=1}^r$ are independent random variables having values in Z a separable metric space, and if g_k is a Borel measurable function, then $\{g_k(\boldsymbol{X}_k)\}_{k=1}^r$ is also independent. Furthermore, if the random variables have values in $\mathbb{R}$, and they are all bounded, then*

$$E\left(\prod_{i=1}^r X_i\right) = \prod_{i=1}^r E(X_i).$$

More generally, the above formula holds if it is only known that each $X_i \in L^1(\Omega;\mathbb{R})$ and

$$\prod_{i=1}^r X_i \in L^1(\Omega;\mathbb{R}).$$

Proof: First consider the claim about $\{g_k(\boldsymbol{X}_k)\}_{k=1}^r$. Letting O be an open set in Z,

$$(g_k \circ \boldsymbol{X}_k)^{-1}(O) = \boldsymbol{X}_k^{-1}\left(g_k^{-1}(O)\right) = \boldsymbol{X}_k^{-1}(\text{Borel set}) \in \sigma(\boldsymbol{X}_k).$$

It follows $(g_k \circ \boldsymbol{X}_k)^{-1}(E)$ is in $\sigma(\boldsymbol{X}_k)$ whenever E is Borel because the sets whose inverse images are measurable includes the Borel sets. Thus $\sigma(g_k \circ \boldsymbol{X}_k) \subseteq \sigma(\boldsymbol{X}_k)$ and this proves the first part of the lemma.

Let $X_1 = \sum_{i=1}^m c_i \mathscr{X}_{E_i}, X_2 = \sum_{j=1}^m d_j \mathscr{X}_{F_j}$ where $P(E_i F_j) = P(E_i) P(F_j)$. Then

$$\int X_1 X_2 dP = \sum_{i,j} d_j c_i P(E_i) P(F_j) = \left(\int X_1 dP\right)\left(\int X_2 dP\right)$$

In general for X_1, X_2 independent, there exist sequences of bounded simple functions, measurable with respect to $\sigma(X_1)$ and $\sigma(X_2)$ respectively $\{s_n\}, \{t_n\}$ such that $s_n \to X_1$ pointwise and $t_n \to X_2$ pointwise, and $|s_n| \leq |X_1|, |t_n| \leq |X_2|$. Then from the above and the dominated convergence theorem,

$$\begin{aligned} \int X_1 X_2 dP &= \lim_{n\to\infty} \int s_n t_n dP = \lim_{n\to\infty} \left(\int s_n dP \right) \left(\int t_n dP \right) \\ &= \left(\int X_1 dP \right) \left(\int X_2 dP \right) \end{aligned}$$

Next suppose there are m of these independent bounded random variables. Then $\prod_{i=2}^m X_i \in \sigma(X_2, \cdots, X_m)$ and by Lemma 26.4.4 the two random variables X_1 and $\prod_{i=2}^m X_i$ are independent. Hence from the above and induction,

$$\int \prod_{i=1}^m X_i dP = \int X_1 \prod_{i=2}^m X_i dP = \int X_1 dP \int \prod_{i=2}^m X_i dP = \prod_{i=1}^m \int X_i dP$$

Now consider the last claim. Replace each X_i with X_i^n where this is just a truncation of the form

$$X_i^n \equiv \begin{cases} X_i \text{ if } |X_i| \leq n \\ n \text{ if } X_i > n \\ -n \text{ if } X_i < n \end{cases}$$

Then by the first part $E\left(\prod_{i=1}^r X_i^n\right) = \prod_{i=1}^r E\left(X_i^n\right)$. Now $\left|\prod_{i=1}^r X_i^n\right| \leq \left|\prod_{i=1}^r X_i\right| \in L^1$ and so by the dominated convergence theorem, you can pass to the limit in both sides to get the desired result. ■

Maybe this would be a good place to put a really interesting result known as the Doob Dynkin lemma. This amazing result is illustrated with the following diagram in which $\boldsymbol{X} = (X_1, \cdots, X_m)$. By Proposition 26.1.6 $\sigma(\boldsymbol{X}) = \sigma(X_1, \cdots, X_n)$.

$$\begin{array}{ccc} (\Omega, \sigma(\boldsymbol{X})) & \xrightarrow{X} & F \\ & \searrow^{\boldsymbol{X}} \quad \nearrow_{g} & \\ & \left(\prod_{i=1}^m E_i, \mathscr{B}\left(\prod_{i=1}^m E_i\right)\right) & \end{array}$$

You start with X and can write it as the composition $g \circ \boldsymbol{X}$ provided X is $\sigma(\boldsymbol{X})$ measurable.

Lemma 26.4.7 *Let $(\Omega, \mathscr{F})$ be a measure space and let $X_i : \Omega \to E_i$ where E_i is a separable Banach space. Suppose also that $X : \Omega \to F$ where F is a separable Banach space. Then X is $\sigma(X_1, \cdots, X_m)$ measurable if and only if there exists a Borel measurable function $g : \prod_{i=1}^m E_i \to F$ such that $X = g(X_1, \cdots, X_m)$.*

Proof: First suppose $X(\omega) = f \mathscr{X}_W(\omega)$ where $f \in F$ and $W \in \sigma(X_1, \cdots, X_m)$. Then by Proposition 26.1.6, W is of the form $(X_1, \cdots, X_m)^{-1}(B) \equiv \boldsymbol{X}^{-1}(B)$ where B is Borel in $\prod_{i=1}^m E_i$. Therefore,

$$X(\omega) = f \mathscr{X}_{\boldsymbol{X}^{-1}(B)}(\omega) = f \mathscr{X}_B(\boldsymbol{X}(\omega)).$$

Now suppose X is measurable with respect to $\sigma(X_1, \cdots, X_m)$. Then there exist simple functions

$$X_n(\omega) = \sum_{k=1}^{m_n} f_k \mathscr{X}_{B_k}(\boldsymbol{X}(\omega)) \equiv g_n(\boldsymbol{X}(\omega))$$

where the B_k are Borel sets in $\prod_{i=1}^m E_i$, such that $X_n(\omega) \to X(\omega)$, each g_n being Borel. Thus g_n converges on $\boldsymbol{X}(\Omega)$. Furthermore, the set on which g_n does converge is a Borel set equal to

$$\cap_{n=1}^{\infty} \cup_{m=1}^{\infty} \cap_{p,q \geq m} \left[||g_p - g_q|| < \frac{1}{n} \right]$$

which contains $\boldsymbol{X}(\Omega)$. Therefore, modifying g_n by multiplying it by the indicator function of this Borel set containing $\boldsymbol{X}(\Omega)$, we can conclude that g_n converges to a Borel function g and, passing to a limit in the above, $X(\omega) = g(\boldsymbol{X}(\omega))$

Conversely, suppose $X(\omega) = g(\boldsymbol{X}(\omega))$. Why is X $\sigma(\boldsymbol{X})$ measurable?

$$X^{-1}(\text{open}) = \boldsymbol{X}^{-1}\left(g^{-1}(\text{open})\right) = \boldsymbol{X}^{-1}(\text{Borel}) \in \sigma(\boldsymbol{X}) \ \blacksquare$$

26.5 Banach Space Valued Random Variables

Recall that for X a random variable, $\sigma(X)$ is the smallest σ algebra containing all the sets of the form $X^{-1}(F)$ where F is Borel. Since such sets, $X^{-1}(F)$ for F Borel form a σ algebra it follows $\sigma(X) = \left\{X^{-1}(F) : F \text{ is Borel}\right\}$.

Next consider the case where you have a set of σ algebras. The following lemma is helpful when you try to verify such a set of σ algebras is independent. It says you only need to check things on π systems contained in the σ algebras. This is really nice because it is much easier to consider the smaller π systems than the whole σ algebra.

Lemma 26.5.1 *Suppose $\{\mathscr{F}_i\}_{i\in I}$ is a set of σ algebras contained in $\mathscr{F}$ where $\mathscr{F}$ is a σ algebra of sets of Ω. Suppose that $\mathscr{K}_i \subseteq \mathscr{F}_i$ is a π system and $\mathscr{F}_i = \sigma(\mathscr{K}_i)$. Suppose also that whenever J is a finite subset of I and $A_j \in \mathscr{K}_j$ for $j \in J$, it follows $P(\cap_{j\in J} A_j) = \prod_{j\in J} P(A_j)$. Then $\{\mathscr{F}_i\}_{i\in I}$ is independent.*

Proof: I need to verify that under the given conditions, if $\{j_1, j_2, \cdots, j_n\} \subseteq I$ and $A_{j_k} \subseteq \mathscr{F}_{j_k}$, then $P\left(\cap_{k=1}^n A_{j_k}\right) = \prod_{k=1}^n P\left(A_{j_k}\right)$. By hypothesis, this is true if each $A_{j_k} \in \mathscr{K}_{j_k}$. Suppose it is true whenever there are at most $r-1 \geq 0$ of the A_{j_k} which are **not** in $\mathscr{K}_{j_k}$. Consider $\cap_{k=1}^n A_{j_k}$ where there are r sets which are **not** in the corresponding $\mathscr{K}_{j_k}$. Without loss of generality, say there are at most $r-1$ sets in the first $n-1$ which are not in the corresponding $\mathscr{K}_{j_k}$.

Pick $\left(A_{j_1} \cdots, A_{j_{n-1}}\right)$ let

$$\mathscr{G}_{\left(A_{j_1} \cdots A_{j_{n-1}}\right)} \equiv \left\{ B \in \mathscr{F}_{j_n} : P\left(\cap_{k=1}^{n-1} A_{j_k} \cap B\right) = \prod_{k=1}^{n-1} P\left(A_{j_k}\right) P(B) \right\}$$

I am going to show $\mathscr{G}_{\left(A_{j_1} \cdots A_{j_{n-1}}\right)}$ is closed with respect to complements and countable disjoint unions and then apply the Lemma on π systems. By the induction hypothesis, $\mathscr{K}_{j_n} \subseteq \mathscr{G}_{\left(A_{j_1} \cdots A_{j_{n-1}}\right)}$. If $B \in \mathscr{G}_{\left(A_{j_1} \cdots A_{j_{n-1}}\right)}$,

$$\prod_{k=1}^{n-1} P\left(A_{j_k}\right) = P\left(\cap_{k=1}^{n-1} A_{j_k}\right) = P\left(\left(\cap_{k=1}^{n-1} A_{j_k} \cap B^C\right) \cup \left(\cap_{k=1}^{n-1} A_{j_k} \cap B\right)\right)$$

$$= P\left(\cap_{k=1}^{n-1} A_{j_k} \cap B^C\right) + P\left(\cap_{k=1}^{n-1} A_{j_k} \cap B\right) = P\left(\cap_{k=1}^{n-1} A_{j_k} \cap B^C\right) + \prod_{k=1}^{n-1} P\left(A_{j_k}\right) P(B)$$

and so

$$P\left(\cap_{k=1}^{n-1}A_{j_k}\cap B^C\right)=\prod_{k=1}^{n-1}P\left(A_{j_k}\right)\left(1-P\left(B\right)\right)=\prod_{k=1}^{n-1}P\left(A_{j_k}\right)P\left(B^C\right)$$

showing if $B\in\mathscr{G}_{\left(A_{j_1}\cdots,A_{j_{n-1}}\right)}$, then so is B^C. It is clear that $\mathscr{G}_{\left(A_{j_1}\cdots,A_{j_{n-1}}\right)}$ is closed with respect to disjoint unions also. Here is why. If $\left\{B_j\right\}_{j=1}^{\infty}$ are disjoint sets in $\mathscr{G}_{\left(A_{j_1}\cdots A_{j_{n-1}}\right)}$,

$$P\left(\cup_{i=1}^{\infty}B_i\cap\cap_{k=1}^{n-1}A_{j_k}\right)=\sum_{i=1}^{\infty}P\left(B_i\cap\cap_{k=1}^{n-1}A_{j_k}\right)=\sum_{i=1}^{\infty}P\left(B_i\right)\prod_{k=1}^{n-1}P\left(A_{j_k}\right)$$

$$=\prod_{k=1}^{n-1}P\left(A_{j_k}\right)\sum_{i=1}^{\infty}P\left(B_i\right)=\prod_{k=1}^{n-1}P\left(A_{j_k}\right)P\left(\cup_{i=1}^{\infty}B_i\right)$$

Therefore, by the π system lemma, Lemma 9.3.2 $\mathscr{G}_{\left(A_{j_1}\cdots A_{j_{n-1}}\right)}=\mathscr{F}_{j_n}$. This proves the induction step in going from $r-1$ to r. ■

What is a useful π system for $\mathscr{B}\left(E\right)$, the Borel sets of E where E is a Banach space?

Recall the fundamental lemma used to prove the Pettis theorem. It was proved on Page 671.

Lemma 26.5.2 *Let E be a separable real Banach space. Sets of the form*

$$\left\{x\in E:x_i^*\left(x\right)\leq\alpha_i,i=1,2,\cdots,m\right\}$$

where $x_i^*\in D'$, a dense subspace of the unit ball of E' and $\alpha_i\in[-\infty,\infty)$ are a π system, and denoting this π system by $\mathscr{K}$, it follows $\sigma\left(\mathscr{K}\right)=\mathscr{B}\left(E\right)$. The sets of $\mathscr{K}$ are examples of cylindrical sets. The D' is that set for the proof of the Pettis theorem.

Proof: The sets described are obviously a π system. I want to show $\sigma\left(\mathscr{K}\right)$ contains the closed balls because then $\sigma\left(\mathscr{K}\right)$ contains the open balls and hence the open sets and the result will follow. Let D' be described in Lemma 24.1.7. As pointed out earlier it can be any dense subset of B'. Then

$$\left\{x\in E:\left\|x-a\right\|\leq r\right\}=\left\{x\in E:\sup_{f\in D'}\left|f\left(x-a\right)\right|\leq r\right\}$$

$$\begin{aligned}&=\left\{x\in E:\sup_{f\in D'}\left|f\left(x\right)-f\left(a\right)\right|\leq r\right\}\\&=\cap_{f\in D'}\left\{x\in E:f\left(a\right)-r\leq f\left(x\right)\leq f\left(a\right)+r\right\}\end{aligned}$$

$$=\cap_{f\in D'}\left\{x\in E:f\left(x\right)\leq f\left(a\right)+r\text{ and }\left(-f\right)\left(x\right)\leq r-f\left(a\right)\right\}$$

which equals a countable intersection of sets of the given π system. Therefore, every closed ball is contained in $\sigma\left(\mathscr{K}\right)$. It follows easily that every open ball is also contained in $\sigma\left(\mathscr{K}\right)$ because

$$B\left(a,r\right)=\cup_{n=1}^{\infty}\overline{B\left(a,r-\frac{1}{n}\right)}.$$

Since the Banach space is separable, it is completely separable and so every open set is the countable union of balls. This shows the open sets are in $\sigma(\mathscr{K})$ and so $\sigma(\mathscr{K}) \supseteq \mathscr{B}(E)$. However, all the sets in the π system are closed hence Borel because they are inverse images of closed sets. Therefore, $\sigma(\mathscr{K}) \subseteq \mathscr{B}(E)$ and so $\sigma(\mathscr{K}) = \mathscr{B}(E)$. ■

As mentioned above, we can replace D' in the above with M, any dense subset of E'.

Observation 26.5.3 *Denote by $C_{\boldsymbol{\alpha},n}$ the set $\{\boldsymbol{\beta} \in \mathbb{R}^n : \beta_i \leq \alpha_i\}$. Also denote by $\boldsymbol{g}_n$ an element of M^n where M is a dense subset of E' with the understanding that $\boldsymbol{g}_n : E \to \mathbb{R}^n$ according to the rule*

$$\boldsymbol{g}_n(x) \equiv (g_1(x), \cdots, g_n(x)).$$

Then the sets in the above lemma can be written as $\boldsymbol{g}_n^{-1}(C_{\boldsymbol{\alpha},n})$. In other words, sets of the form $\boldsymbol{g}_n^{-1}(C_{\boldsymbol{\alpha},n})$ form a π system for $\mathscr{B}(E)$.

Next suppose you have some random variables having values in a separable Banach space, E, $\{X_i\}_{i\in I}$. How can you tell if they are independent? To show they are independent, you need to verify that

$$P\left(\cap_{k=1}^n X_{i_k}^{-1}(F_{i_k})\right) = \prod_{k=1}^n P\left(X_{i_k}^{-1}(F_{i_k})\right)$$

whenever the F_{i_k} are Borel sets in E. It is desirable to find a way to do this easily.

Lemma 26.5.4 *Let $\mathscr{K}$ be a π system of sets of E, a separable real Banach space and let $(\Omega, \mathscr{F}, P)$ be a probability space and $X : \Omega \to E$ be a random variable. Then*

$$X^{-1}(\sigma(\mathscr{K})) = \sigma\left(X^{-1}(\mathscr{K})\right)$$

Proof: First note that $X^{-1}(\sigma(\mathscr{K}))$ is a σ algebra which contains $X^{-1}(\mathscr{K})$ and so it contains $\sigma\left(X^{-1}(\mathscr{K})\right)$. Thus $X^{-1}(\sigma(\mathscr{K})) \supseteq \sigma\left(X^{-1}(\mathscr{K})\right)$. Now let

$$\mathscr{G} \equiv \left\{A \in \sigma(\mathscr{K}) : X^{-1}(A) \in \sigma\left(X^{-1}(\mathscr{K})\right)\right\}$$

Then $\mathscr{G} \supseteq \mathscr{K}$. If $A \in \mathscr{G}$, then $X^{-1}(A) \in \sigma\left(X^{-1}(\mathscr{K})\right)$ and so

$$X^{-1}(A)^C = X^{-1}\left(A^C\right) \in \sigma\left(X^{-1}(\mathscr{K})\right)$$

because $\sigma\left(X^{-1}(\mathscr{K})\right)$ is a σ algebra. Hence $A^C \in \mathscr{G}$. Finally suppose $\{A_i\}$ is a sequence of disjoint sets of $\mathscr{G}$. Then

$$X^{-1}(\cup_{i=1}^\infty A_i) = \cup_{i=1}^\infty X^{-1}(A_i) \in \sigma\left(X^{-1}(\mathscr{K})\right)$$

again because $\sigma\left(X^{-1}(\mathscr{K})\right)$ is a σ algebra. It follows from Lemma 9.3.2 on Page 247 that $\mathscr{G} \supseteq \sigma(\mathscr{K})$ and this shows that whenever

$$A \in \sigma(\mathscr{K}),\ X^{-1}(A) \in \sigma\left(X^{-1}(\mathscr{K})\right).$$

Thus $X^{-1}(\sigma(\mathscr{K})) \subseteq \sigma\left(X^{-1}(\mathscr{K})\right)$. ■

With this lemma, here is the desired result about independent random variables. Essentially, you can reduce to the case of random vectors having values in $\mathbb{R}^n$.

26.6 Reduction to Finite Dimensions

Let E be a Banach space and let $\boldsymbol{g} \in (E')^n$. Then for $x \in E$, $\boldsymbol{g} \circ x$ is the vector in $\mathbb{F}^n$ which equals $(g_1(x), g_2(x), \cdots, g_n(x))$.

Theorem 26.6.1 *Let X_i be a random variable having values in E a real separable Banach space. The random variables $\{X_i\}_{i\in I}$ are independent if whenever*

$$\{i_1, \cdots, i_n\} \subseteq I,$$

$m_{i_1}, \cdots, m_{i_n}$ are positive integers, and $\boldsymbol{g}_{m_{i_1}}, \cdots, \boldsymbol{g}_{m_{i_n}}$ are respectively in

$$(M)^{m_{i_1}}, \cdots, (M)^{m_{i_n}}$$

for M a dense subspace of E', $\left\{\boldsymbol{g}_{m_{i_j}} \circ X_{i_j}\right\}_{j=1}^n$ are independent random vectors having values in $\mathbb{R}^{m_{i_1}}, \cdots, \mathbb{R}^{m_{i_n}}$ respectively.

Proof: It is necessary to show that the events $X_{i_j}^{-1}\left(B_{i_j}\right)$ are independent events whenever B_{i_j} are Borel sets. By Lemma 26.5.1 and the above Lemma 26.5.2, it suffices to verify that the events

$$X_{i_j}^{-1}\left(\boldsymbol{g}_{m_{i_j}}^{-1}\left(C_{\vec{\alpha},m_{i_j}}\right)\right) = \left(\boldsymbol{g}_{m_{i_j}} \circ X_{i_j}\right)^{-1}\left(C_{\vec{\alpha},m_{i_j}}\right)$$

are independent where $C_{\vec{\alpha},m_{i_j}}$ are the cones described in Lemma 26.5.2. Thus

$$\vec{\alpha} = \left(\alpha_{k_1}, \cdots, \alpha_{k_m}\right),\ C_{\vec{\alpha},m_{i_j}} = \prod_{i=1}^{m_{i_j}} (-\infty, \alpha_{k_i}]$$

But this condition is implied when the finite dimensional valued random vectors $\boldsymbol{g}_{m_{i_j}} \circ X_{i_j}$ are independent. ■

The above assertion also goes the other way as you may want to show.

26.7 $0,1$ Laws

I am following [54] for the proof of many of the following theorems. Recall the set of ω which are in infinitely many of the sets $\{A_n\}$ is $\cap_{n=1}^\infty \cup_{m=n}^\infty A_m$. This is in $\cap_{n=1}^\infty \cup_{m=n}^\infty A_m$ if and only if for every n there exists $m \geq n$ such that it is in A_m.

Theorem 26.7.1 *Suppose $A_n \in \mathcal{F}_n$ where the σ algebras $\{\mathcal{F}_n\}_{n=1}^\infty$ are independent. Suppose also that $\sum_{k=1}^\infty P(A_k) = \infty$. Then $P\left(\cap_{n=1}^\infty \cup_{m=n}^\infty A_m\right) = 1$.*

Proof: It suffices to verify that $P\left(\cup_{n=1}^\infty \cap_{m=n}^\infty A_m^C\right) = 0$ which can be accomplished by showing that $P\left(\cap_{m=n}^\infty A_m^C\right) = 0$ for each n. The sets $\left\{A_k^C\right\}$ satisfy $A_k^C \in \mathcal{F}_k$. Therefore, noting that $e^{-x} \geq 1 - x$,

$$\begin{aligned} P\left(\cap_{m=n}^\infty A_m^C\right) &= \lim_{N\to\infty} P\left(\cap_{m=n}^N A_m^C\right) = \lim_{N\to\infty} \prod_{m=n}^N P\left(A_m^C\right) \\ &= \lim_{N\to\infty} \prod_{m=n}^N (1 - P(A_m)) \leq \lim_{N\to\infty} \prod_{m=n}^N e^{-P(A_m)} \\ &= \lim_{N\to\infty} \exp\left(-\sum_{m=n}^N P(A_m)\right) = 0. \ ■ \end{aligned}$$

The Kolmogorov zero one law follows next. It has to do with something called a tail event.

Definition 26.7.2 *Let $\{\mathcal{F}_n\}$ be a sequence of σ algebras. Then $\mathcal{T}_n \equiv \sigma\left(\cup_{k=n}^{\infty}\mathcal{F}_k\right)$ where this means the smallest σ algebra which contains each $\mathcal{F}_k$ for $k \geq n$. Then a tail event is a set which is in the σ algebra, $\mathcal{T} \equiv \cap_{n=1}^{\infty}\mathcal{T}_n$.*

As usual, $(\Omega, \mathcal{F}, P)$ is the underlying probability space such that all σ algebras are contained in $\mathcal{F}$.

Lemma 26.7.3 *Suppose $\{\mathcal{F}_n\}_{n=1}^{\infty}$ are independent σ algebras and suppose A is a tail event and $A_{k_i} \in \mathcal{F}_{k_i}$, $i = 1, \cdots, m$ are given sets. Then*

$$P\left(A_{k_1} \cap \cdots \cap A_{k_m} \cap A\right) = P\left(A_{k_1} \cap \cdots \cap A_{k_m}\right) P(A)$$

Proof: Let $\mathcal{K}$ be the π system consisting of finite intersections of the form

$$B_{m_1} \cap B_{m_2} \cap \cdots \cap B_{m_j}$$

where $B_{m_i} \in \mathcal{F}_{k_i}$ for $k_i > \max\{k_1, \cdots, k_m\} \equiv N$. Thus $\sigma(\mathcal{K}) = \sigma\left(\cup_{i=N+1}^{\infty}\mathcal{F}_i\right)$. Now let

$$\mathcal{G} \equiv \left\{B \in \sigma(\mathcal{K}) : P\left(A_{k_1} \cap \cdots \cap A_{k_m} \cap B\right) = P\left(A_{k_1} \cap \cdots \cap A_{k_m}\right) P(B)\right\}$$

Then clearly $\mathcal{K} \subseteq \mathcal{G}$. It is also true that $\mathcal{G}$ is closed with respect to complements and countable disjoint unions. By the lemma on π systems, $\mathcal{G} = \sigma(\mathcal{K}) = \sigma\left(\cup_{i=N+1}^{\infty}\mathcal{F}_i\right)$. Since A is in $\sigma\left(\cup_{i=N+1}^{\infty}\mathcal{F}_i\right)$ due to the assumption that it is a tail event, it follows that

$$P\left(A_{k_1} \cap \cdots \cap A_{k_m} \cap A\right) = P\left(A_{k_1} \cap \cdots \cap A_{k_m}\right) P(A) \ \blacksquare$$

Theorem 26.7.4 *Suppose the σ algebras, $\{\mathcal{F}_n\}_{n=1}^{\infty}$ are independent and suppose A is a tail event. Then $P(A)$ either equals 0 or 1.*

Proof: Let $A \in \mathcal{T} \equiv \cap_{n=1}^{\infty}\mathcal{T}_n \equiv \cap_{n=1}^{\infty}\sigma\left(\cup_{k=n}^{\infty}\mathcal{F}_k\right)$. I want to show that $P(A) = P(A)^2$. Since A is in $\mathcal{T}$, it is in each $\sigma\left(\cup_{k=n}^{\infty}\mathcal{F}_k\right)$. Let $\mathcal{K}$ denote sets of the form $A_{k_1} \cap \cdots \cap A_{k_m}$ for some m, $A_{k_j} \in \mathcal{F}_{k_j}$ where each $k_j > n$. Thus $\mathcal{K}$ is a π system and

$$\sigma(\mathcal{K}) = \sigma\left(\cup_{k=n+1}^{\infty}\mathcal{F}_k\right) \equiv \mathcal{T}_{n+1}$$

Let

$$\mathcal{G} \equiv \left\{B \in \mathcal{T}_{n+1} \equiv \sigma\left(\cup_{k=n+1}^{\infty}\mathcal{F}_k\right) : P(A \cap B) = P(A) P(B)\right\}$$

Thus $\mathcal{K} \subseteq \mathcal{G}$ because

$$P\left(A_{k_1} \cap \cdots \cap A_{k_m} \cap A\right) = P\left(A_{k_1} \cap \cdots \cap A_{k_m}\right) P(A)$$

by Lemma 26.7.3. However, it is routine that $\mathcal{G}$ is closed with respect to countable disjoint unions and complements. Therefore by the Lemma on π systems Lemma 9.3.2 on Page 247, it follows $\mathcal{G} = \sigma(\mathcal{K}) = \sigma\left(\cup_{k=n+1}^{\infty}\mathcal{F}_k\right)$.

Thus for any $B \in \sigma\left(\cup_{k=n+1}^{\infty}\mathcal{F}_k\right) = \mathcal{T}_{n+1}, P(A \cap B) = P(A) P(B)$. However, A is in all of these $\mathcal{T}_{n+1}$ and so $P(A \cap A) = P(A) = P(A)^2$ so $P(A)$ equals either 0 or 1. $\blacksquare$

What sorts of things are tail events of independent σ algebras?

Theorem 26.7.5 *Let $\{\boldsymbol{X}_k\}$ be a sequence of independent random variables having values in Z a Banach space. That is, the σ algebras $\sigma(X_k)$ are independent. Then*

$$A \equiv \{\omega : \{\boldsymbol{X}_k(\omega)\} \text{ converges}\}$$

is a tail event. So is

$$B \equiv \left\{\omega : \left\{\sum_{k=1}^{\infty} \boldsymbol{X}_k(\omega)\right\} \text{ converges}\right\}.$$

Proof: Since Z is complete, A is the same as the set where $\{\boldsymbol{X}_k(\omega)\}$ is a Cauchy sequence. This set is

$$\cap_{n=1}^{\infty}\cap_{p=1}^{\infty}\cup_{m=p}^{\infty}\cap_{l,k\geq m}\{\omega : \|\boldsymbol{X}_k(\omega)-\boldsymbol{X}_l(\omega)\| < 1/n\}$$

Note that

$$\cup_{m=p}^{\infty}\cap_{l,k\geq m}\{\omega : \|\boldsymbol{X}_k(\omega)-\boldsymbol{X}_l(\omega)\| < 1/n\} \in \sigma\left(\cup_{j=p}^{\infty}\sigma(\boldsymbol{X}_j)\right)$$

for every p is the set where ultimately any pair of $\boldsymbol{X}_k, \boldsymbol{X}_l$ are closer together than $1/n$,

$$\cap_{p=1}^{\infty}\cup_{m=p}^{\infty}\cap_{l,k\geq m}\{\omega : \|\boldsymbol{X}_k(\omega)-\boldsymbol{X}_l(\omega)\| < 1/n\}$$

is a tail event. The set where $\{\boldsymbol{X}_k(\omega)\}$ is a Cauchy sequence is the intersection of all these and is therefore, also a tail event.

Now consider B. This set is the same as the set where the partial sums are Cauchy sequences. Let $\boldsymbol{S}_n \equiv \sum_{k=1}^{n}\boldsymbol{X}_k$. The set where the sum converges is then

$$\cap_{n=1}^{\infty}\cap_{p=2}^{\infty}\cup_{m=p}^{\infty}\cap_{l,k\geq m}\{\omega : \|\boldsymbol{S}_k(\omega)-\boldsymbol{S}_l(\omega)\| < 1/n\}$$

Say $k<l$ and consider for $m \geq p$

$$\{\omega : \|\boldsymbol{S}_k(\omega)-\boldsymbol{S}_l(\omega)\| < 1/n,\ k \geq m\}$$

This is the same as

$$\left\{\omega : \left\|\sum_{j=k-1}^{l}\boldsymbol{X}_j(\omega)\right\| < 1/n, k \geq m\right\} \in \sigma\left(\cup_{j=p-1}^{\infty}\sigma(\boldsymbol{X}_j)\right)$$

Thus

$$\cup_{m=p}^{\infty}\cap_{l,k\geq m}\{\omega : \|\boldsymbol{S}_k(\omega)-\boldsymbol{S}_l(\omega)\| < 1/n\} \in \sigma\left(\cup_{j=p-1}^{\infty}\sigma(\boldsymbol{X}_j)\right)$$

and so the intersection for all p of these is a tail event. Then the intersection over all n of these tail events is a tail event. ■

From this it can be concluded that if you have a sequence of independent random variables, $\{\boldsymbol{X}_k\}$ the set where it converges is either of probability 1 or probability 0. A similar conclusion holds for the set where the infinite sum of these random variables converges. This is stated in the next corollary. This incredible assertion is the next corollary.

Corollary 26.7.6 *Let $\{\boldsymbol{X}_k\}$ be a sequence of random variables having values in a Banach space. Then $\lim_{n\to\infty}\boldsymbol{X}_n(\omega)$ either exists for a.e. ω or the convergence fails to take place for a.e. ω. Also if*

$$A \equiv \left\{\omega : \sum_{k=1}^{\infty}\boldsymbol{X}_k(\omega) \text{ converges}\right\},$$

then $P(A) = 0$ or 1.

26.8 Strong Law of Large Numbers

Kolmogorov's inequality is a very interesting inequality which depends on independence of a set of random vectors. The random vectors have values in $\mathbb{R}^n$ or more generally some real separable Hilbert space.

Lemma 26.8.1 *If $\boldsymbol{Y}, \boldsymbol{X}$ are independent random variables having values in a real separable Hilbert space, H with $E\left(|\boldsymbol{X}|^2\right), E\left(|\boldsymbol{Y}|^2\right) < \infty$, then*

$$\int_\Omega (\boldsymbol{X}, \boldsymbol{Y})\, dP = \left(\int_\Omega \boldsymbol{X}\, dP, \int_\Omega \boldsymbol{Y}\, dP \right).$$

Proof: Let $\{\boldsymbol{e}_k\}$ be a complete orthonormal basis. Thus from Theorem 22.4.2,

$$\int_\Omega (\boldsymbol{X}, \boldsymbol{Y})\, dP = \int_\Omega \sum_{k=1}^{\infty} (\boldsymbol{X}, \boldsymbol{e}_k)(\boldsymbol{Y}, \boldsymbol{e}_k)\, dP$$

Now

$$\int_\Omega \sum_{k=1}^{\infty} |(\boldsymbol{X}, \boldsymbol{e}_k)(\boldsymbol{Y}, \boldsymbol{e}_k)|\, dP \leq \int_\Omega \left(\sum_k |(\boldsymbol{X}, \boldsymbol{e}_k)|^2 \right)^{1/2} \left(\sum_k |(\boldsymbol{Y}, \boldsymbol{e}_k)|^2 \right)^{1/2} dP$$

$$= \int_\Omega |\boldsymbol{X}|\, |\boldsymbol{Y}|\, dP \leq \left(\int_\Omega |\boldsymbol{X}|^2\, dP \right)^{1/2} \left(\int_\Omega |\boldsymbol{Y}|^2\, dP \right)^{1/2} < \infty$$

and so by Fubini's theorem and independence of $\boldsymbol{X}, \boldsymbol{Y}$,

$$\int_\Omega (\boldsymbol{X}, \boldsymbol{Y})\, dP = \int_\Omega \sum_{k=1}^{\infty} (\boldsymbol{X}, \boldsymbol{e}_k)(\boldsymbol{Y}, \boldsymbol{e}_k)\, dP = \sum_{k=1}^{\infty} \int_\Omega (\boldsymbol{X}, \boldsymbol{e}_k)(\boldsymbol{Y}, \boldsymbol{e}_k)\, dP$$

$$= \sum_{k=1}^{\infty} \int_\Omega (\boldsymbol{X}, \boldsymbol{e}_k)\, dP \int_\Omega (\boldsymbol{Y}, \boldsymbol{e}_k)\, dP = \sum_{k=1}^{\infty} \left(\int_\Omega \boldsymbol{X}\, dP, \boldsymbol{e}_k \right) \left(\int_\Omega \boldsymbol{Y}\, dP, \boldsymbol{e}_k \right) dP$$

$$= \left(\int_\Omega \boldsymbol{X}\, dP, \int_\Omega \boldsymbol{Y}\, dP \right) \blacksquare$$

Now here is Kolmogorov's inequality.

Theorem 26.8.2 *Suppose $\{\boldsymbol{X}_k\}_{k=1}^n$ are independent with $E(|\boldsymbol{X}_k|) < \infty$, $E(\boldsymbol{X}_k) = \boldsymbol{0}$. Then for any $\varepsilon > 0$,*

$$P\left(\left[\max_{1 \leq k \leq n} \left| \sum_{j=1}^{k} \boldsymbol{X}_j \right| \geq \varepsilon \right] \right) \leq \frac{1}{\varepsilon^2} \sum_{j=1}^{n} E\left(|\boldsymbol{X}_k|^2 \right).$$

Proof: Let $A = \left[\max_{1 \leq k \leq n} \left| \sum_{j=1}^{k} \boldsymbol{X}_j \right| \geq \varepsilon \right]$. Now let $A_1 \equiv [|\boldsymbol{X}_1| \geq \varepsilon]$ and if $A_1, \cdots, A_m$ have been chosen,

$$A_{m+1} \equiv \left[\left| \sum_{j=1}^{m+1} \boldsymbol{X}_j \right| \geq \varepsilon \right] \cap \bigcap_{r=1}^{m} \left[\left| \sum_{j=1}^{r} \boldsymbol{X}_j \right| < \varepsilon \right]$$

Thus the A_k partition A and $\omega \in A_k$ means $\left|\sum_{j=1}^{k} \boldsymbol{X}_j\right| \geq \varepsilon$ but this did not happen for $\left|\sum_{j=1}^{r} \boldsymbol{X}_j\right|$ for any $r < k$. Note also that $A_k \in \sigma(\boldsymbol{X}_1, \cdots, \boldsymbol{X}_k)$. Then from algebra,

$$\left|\sum_{j=1}^{n} \boldsymbol{X}_j\right|^2 = \left(\sum_{i=1}^{k} \boldsymbol{X}_i + \sum_{j=k+1}^{n} \boldsymbol{X}_j, \sum_{i=1}^{k} \boldsymbol{X}_i + \sum_{j=k+1}^{n} \boldsymbol{X}_j\right)$$

$$= \left|\sum_{j=1}^{k} \boldsymbol{X}_j\right|^2 + \sum_{i\leq k, j>k} (\boldsymbol{X}_i, \boldsymbol{X}_j) + \sum_{i\leq k, j>k} (\boldsymbol{X}_j, \boldsymbol{X}_i) + \sum_{i>k, j>k} (\boldsymbol{X}_j, \boldsymbol{X}_i)$$

Written more succinctly, $\left|\sum_{j=1}^{n} \boldsymbol{X}_j\right|^2 = \left|\sum_{j=1}^{k} \boldsymbol{X}_j\right|^2 + \sum_{j>k \text{ or } i>k} (\boldsymbol{X}_i, \boldsymbol{X}_j)$. Now multiply both sides by $\mathcal{X}_{A_k}$ and integrate. Suppose $i \leq k$ for one of the terms in the second sum. Then by Lemma 26.4.4 and $A_k \in \sigma(\boldsymbol{X}_1, \cdots, \boldsymbol{X}_k)$, the two random vectors $\mathcal{X}_{A_k}\boldsymbol{X}_i, \boldsymbol{X}_j$ are independent,

$$\int_\Omega \mathcal{X}_{A_k} (\boldsymbol{X}_i, \boldsymbol{X}_j)\, dP = \left(\int_\Omega \mathcal{X}_{A_k} \boldsymbol{X}_i dP, \int_\Omega \boldsymbol{X}_j dP\right) = 0$$

the last equality holding because by assumption $E(\boldsymbol{X}_j) = \boldsymbol{0}$. Therefore, it can be assumed both i, j are larger than k and

$$\int_\Omega \mathcal{X}_{A_k} \left|\sum_{j=1}^{n} \boldsymbol{X}_j\right|^2 dP = \int_\Omega \mathcal{X}_{A_k} \left|\sum_{j=1}^{k} \boldsymbol{X}_j\right|^2 dP + \sum_{j>k, i>k} \int_\Omega \mathcal{X}_{A_k} (\boldsymbol{X}_i, \boldsymbol{X}_j)\, dP \tag{26.8}$$

The last term on the right is interesting. Suppose $i > j$. The integral inside the sum is of the form $\int_\Omega \left(\boldsymbol{X}_i, \mathcal{X}_{A_k}\boldsymbol{X}_j\right) dP$. The second factor in the inner product is in

$$\sigma(\boldsymbol{X}_1, \cdots, \boldsymbol{X}_k, \boldsymbol{X}_j)$$

and $\boldsymbol{X}_i$ is not included in the list of random vectors. Thus by Lemma 26.4.4, the two random vectors $\boldsymbol{X}_i, \mathcal{X}_{A_k}\boldsymbol{X}_j$ are independent and so the last term in 26.8 reduces to

$$\left(\int_\Omega \boldsymbol{X}_i dP, \int_\Omega \mathcal{X}_{A_k}\boldsymbol{X}_j dP\right) = \left(\boldsymbol{0}, \int_\Omega \mathcal{X}_{A_k}\boldsymbol{X}_j dP\right) = 0.$$

A similar result holds if $j > i$. Thus the mixed terms in the last term of 26.8 are all equal to 0. Hence 26.8 reduces to

$$\int_\Omega \mathcal{X}_{A_k} \left|\sum_{j=1}^{n} \boldsymbol{X}_j\right|^2 dP = \int_\Omega \mathcal{X}_{A_k} \left|\sum_{j=1}^{k} \boldsymbol{X}_j\right|^2 dP + \sum_{i>k} \int_\Omega \mathcal{X}_{A_k} |\boldsymbol{X}_i|^2 dP$$

and so $\int_\Omega \mathcal{X}_{A_k} \left|\sum_{j=1}^{n} \boldsymbol{X}_j\right|^2 dP \geq \int_\Omega \mathcal{X}_{A_k} \left|\sum_{j=1}^{k} \boldsymbol{X}_j\right|^2 dP \geq \varepsilon^2 P(A_k)$ .Now, summing these yields

$$\varepsilon^2 P(A) \leq \int_\Omega \mathcal{X}_A \left|\sum_{j=1}^{n} \boldsymbol{X}_j\right|^2 dP \leq \int_\Omega \left|\sum_{j=1}^{n} \boldsymbol{X}_j\right|^2 dP = \sum_{i,j} \int_\Omega (\boldsymbol{X}_i, \boldsymbol{X}_j)\, dP$$

By independence of the random vectors the mixed terms of the above sum equal zero and so it reduces to $\sum_{i=1}^{n} \int_\Omega |\boldsymbol{X}_i|^2 dP$ ■

This theorem implies the following amazing result.

Theorem 26.8.3 *Let $\{\boldsymbol{X}_k\}_{k=1}^{\infty}$ be independent random vectors having values in a separable real Hilbert space and suppose $E(|\boldsymbol{X}_k|) < \infty$ for each k and $E(\boldsymbol{X}_k) = \mathbf{0}$. Suppose also that $\sum_{j=1}^{\infty} E\left(|\boldsymbol{X}_j|^2\right) < \infty$. Then $\sum_{j=1}^{\infty} \boldsymbol{X}_j$ converges a.e.*

Proof: Let $\varepsilon > 0$ be given. By Kolmogorov's inequality, Theorem 26.8.2, it follows that for $p \leq m < n$

$$P\left(\left[\max_{m\leq k\leq n}\left|\sum_{j=m}^{k}\boldsymbol{X}_j\right| \geq \varepsilon\right]\right) \leq \frac{1}{\varepsilon^2}\sum_{j=p}^{n} E\left(|\boldsymbol{X}_j|^2\right) \leq \frac{1}{\varepsilon^2}\sum_{j=p}^{\infty} E\left(|\boldsymbol{X}_j|^2\right).$$

Therefore, letting $n \to \infty$ it follows that for all m,n such that $p \leq m \leq n$

$$P\left(\left[\max_{p\leq m\leq n}\left|\sum_{j=m}^{n}\boldsymbol{X}_j\right| \geq \varepsilon\right]\right) \leq \frac{1}{\varepsilon^2}\sum_{j=p}^{\infty} E\left(|\boldsymbol{X}_j|^2\right).$$

It follows from the assumption $\sum_{j=1}^{\infty} E\left(|\boldsymbol{X}_j|^2\right) < \infty$ there exists a sequence, $\{p_n\}$ such that if $m \geq p_n$

$$P\left(\left[\max_{k\geq m\geq p_n}\left|\sum_{j=m}^{k}\boldsymbol{X}_j\right| \geq 2^{-n}\right]\right) \leq 2^{-n}.$$

By the Borel Cantelli lemma, Lemma 26.1.2, there is a set of measure 0, N such that for $\omega \notin N$, ω is in only finitely many of the sets,

$$\left[\max_{k\geq m\geq p_n}\left|\sum_{j=m}^{k}\boldsymbol{X}_j\right| \geq 2^{-n}\right]$$

and so for $\omega \notin N$, it follows that for large enough n,

$$\left[\max_{k\geq m\geq p_n}\left|\sum_{j=m}^{k}\boldsymbol{X}_j(\omega)\right| < 2^{-n}\right].$$

However, this says the partial sums $\left\{\sum_{j=1}^{k}\boldsymbol{X}_j(\omega)\right\}_{k=1}^{\infty}$ are a Cauchy sequence. Therefore, they converge. ■

With this amazing result, there is a simple proof of the strong law of large numbers but first is an elementary lemma. In the following lemma, s_k and a_j could have values in any normed linear space.

Lemma 26.8.4 *Suppose $s_k \to s$. Then $\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n} s_k = s$. Also if $\sum_{j=1}^{\infty}\frac{a_j}{j}$ converges, then $\lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n} a_j = 0$.*

Proof: Consider the first part. Since $s_k \to s$, it follows there is some constant, C such that $|s_k| < C$ for all k and $|s| < C$ also. Choose K so large that if $k \geq K$, then for $n > K$,

$$|s - s_k| < \varepsilon/2.$$

$$\left|s - \frac{1}{n}\sum_{k=1}^{n} s_k\right| \leq \frac{1}{n}\sum_{k=1}^{n}|s_k - s| = \frac{1}{n}\sum_{k=1}^{K}|s_k - s| + \frac{1}{n}\sum_{k=K}^{n}|s_k - s|$$

$$\leq \frac{2CK}{n} + \frac{\varepsilon}{2}\frac{n-K}{n} < \frac{2CK}{n} + \frac{\varepsilon}{2}$$

Therefore, whenever n is large enough, $\left|s - \frac{1}{n}\sum_{k=1}^{n} s_k\right| < \varepsilon$.

Now consider the second claim. Let $s_k = \sum_{j=1}^{k} \frac{a_j}{j}$ and $s = \lim_{k\to\infty} s_k$ Then by the first part,

$$\begin{aligned} s &= \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} s_k = \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n}\sum_{j=1}^{k} \frac{a_j}{j} \\ &= \lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} \frac{a_j}{j}\sum_{k=j}^{n} 1 = \lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} \frac{a_j}{j}(n-j) \\ &= \lim_{n\to\infty}\left(\sum_{j=1}^{n} \frac{a_j}{j} - \frac{1}{n}\sum_{j=1}^{n} a_j\right) = s - \lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} a_j \ \blacksquare \end{aligned}$$

Now here is the strong law of large numbers.

Theorem 26.8.5 *Suppose $\{\boldsymbol{X}_k\}$ are independent random variables, and also suppose that $E(|\boldsymbol{X}_k|) < \infty$ for each k and $E(\boldsymbol{X}_k) = \boldsymbol{m}_k$. Suppose also*

$$\sum_{j=1}^{\infty} \frac{1}{j^2} E\left(\left|\boldsymbol{X}_j - \boldsymbol{m}_j\right|^2\right) < \infty. \tag{26.9}$$

Then $\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} (\boldsymbol{X}_j - \boldsymbol{m}_j) = \boldsymbol{0}$.

Proof: Consider the sum $\sum_{j=1}^{\infty} \frac{\boldsymbol{X}_j - \boldsymbol{m}_j}{j}$. This sum converges *a.e.* because of 26.9 and Theorem 26.8.3 applied to the random vectors $\left\{\frac{\boldsymbol{X}_j - \boldsymbol{m}_j}{j}\right\}$. Therefore, from Lemma 26.8.4 it follows that for a.e. ω, $\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} (\boldsymbol{X}_j(\omega) - \boldsymbol{m}_j) = \boldsymbol{0}$ $\blacksquare$

The next corollary is often called the strong law of large numbers. It follows immediately from the above theorem.

Corollary 26.8.6 *Suppose $\{\boldsymbol{X}_j\}_{j=1}^{\infty}$ are independent having mean $\boldsymbol{m}$ and variance equal to $\sigma^2 \equiv \int_\Omega \left|\boldsymbol{X}_j - \boldsymbol{m}\right|^2 dP < \infty$. Then for a.e. $\omega \in \Omega$,*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{X}_j(\omega) = \boldsymbol{m}$$

Chapter 27

Analytical Considerations

27.1 The Characteristic Function

One of the most important tools in probability is the characteristic function. To begin with, assume the random variables have values in $\mathbb{R}^p$.

Definition 27.1.1 *Let $\boldsymbol{X}$ be a random variable as above. The characteristic function is*

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) \equiv E\left(e^{i\boldsymbol{t}\cdot\boldsymbol{X}}\right) \equiv \int_\Omega e^{i\boldsymbol{t}\cdot\boldsymbol{X}(\omega)}dP = \int_{\mathbb{R}^p} e^{i\boldsymbol{t}\cdot\boldsymbol{x}}d\lambda_{\boldsymbol{X}}$$

the last equation holding by Proposition 26.1.12.

Recall the following fundamental lemma and definition, Lemma 13.2.4 on Page 387 where $\mathcal{G}$ was a set of functions described there.

Definition 27.1.2 *For $T \in \mathcal{G}^*$, define $FT, F^{-1}T \in \mathcal{G}^*$ by*

$$FT(\phi) \equiv T(F\phi),\ F^{-1}T(\phi) \equiv T\left(F^{-1}\phi\right)$$

Lemma 27.1.3 *F and F^{-1} are both one to one, onto, and are inverses of each other.*

The main result on characteristic functions is the following.

Theorem 27.1.4 *Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be random vectors with values in $\mathbb{R}^p$ and suppose $E\left(e^{i\boldsymbol{t}\cdot\boldsymbol{X}}\right) = E\left(e^{i\boldsymbol{t}\cdot\boldsymbol{Y}}\right)$ for all $\boldsymbol{t} \in \mathbb{R}^p$. Then $\lambda_{\boldsymbol{X}} = \lambda_{\boldsymbol{Y}}$.*

Proof: For $\psi \in \mathcal{G}$, let $\lambda_{\boldsymbol{X}}(\psi) \equiv \int_{\mathbb{R}^p} \psi d\lambda_{\boldsymbol{X}}$ and $\lambda_{\boldsymbol{Y}}(\psi) \equiv \int_{\mathbb{R}^p} \psi d\lambda_{\boldsymbol{Y}}$. Thus both $\lambda_{\boldsymbol{X}}$ and $\lambda_{\boldsymbol{Y}}$ are in $\mathcal{G}^*$. Then letting $\psi \in \mathcal{G}$ and using Fubini's theorem,

$$\begin{aligned}
\int_{\mathbb{R}^p}\int_{\mathbb{R}^p} e^{i\boldsymbol{t}\cdot\boldsymbol{y}}\psi(\boldsymbol{t})\,dt\,d\lambda_{\boldsymbol{Y}} &= \int_{\mathbb{R}^p}\int_{\mathbb{R}^p} e^{i\boldsymbol{t}\cdot\boldsymbol{y}}d\lambda_{\boldsymbol{Y}}\psi(\boldsymbol{t})\,dt = \int_{\mathbb{R}^p} E\left(e^{i\boldsymbol{t}\cdot\boldsymbol{Y}}\right)\psi(\boldsymbol{t})\,dt \\
&= \int_{\mathbb{R}^p} E\left(e^{i\boldsymbol{t}\cdot\boldsymbol{X}}\right)\psi(\boldsymbol{t})\,dt = \int_{\mathbb{R}^p}\int_{\mathbb{R}^p} e^{i\boldsymbol{t}\cdot\boldsymbol{x}}d\lambda_{\boldsymbol{X}}\psi(\boldsymbol{t})\,dt \\
&= \int_{\mathbb{R}^p}\int_{\mathbb{R}^p} e^{i\boldsymbol{t}\cdot\boldsymbol{x}}\psi(\boldsymbol{t})\,dt\,d\lambda_{\boldsymbol{X}}.
\end{aligned}$$

Thus $\lambda_{\boldsymbol{Y}}\left(F^{-1}\psi\right) = \lambda_{\boldsymbol{X}}\left(F^{-1}\psi\right)$. Since $\psi \in \mathcal{G}$ is arbitrary and F^{-1} is onto, this implies $\lambda_{\boldsymbol{X}} = \lambda_{\boldsymbol{Y}}$ in $\mathcal{G}^*$. But $\mathcal{G}$ is dense in $C_0(\mathbb{R}^p)$ from the Stone Weierstrass theorem and so $\lambda_{\boldsymbol{X}} = \lambda_{\boldsymbol{Y}}$ as measures. Recall from real analysis the dual space of $C_0(\mathbb{R}^n)$ is the space of complex measures.

Alternatively, the above shows that since F^{-1} is onto, for all $\psi \in \mathcal{G}$,

$$\int_{\mathbb{R}^p} \psi d\lambda_{\boldsymbol{Y}} = \int_{\mathbb{R}^p} \psi d\lambda_{\boldsymbol{X}}$$

and then, by a use of the Stone Weierstrass theorem, the above will hold for all $\psi \in C_c(\mathbb{R}^n)$ and now, by the Riesz representation theorem for positive linear functionals, $\lambda_{\boldsymbol{Y}} = \lambda_{\boldsymbol{X}}$. ■

You can also give a version of this theorem in which reference is made only to the probability distribution measures.

Definition 27.1.5 *For μ a probability measure on the Borel sets of $\mathbb{R}^n$,*

$$\phi_\mu(t) \equiv \int_{\mathbb{R}^n} e^{it\cdot x} d\mu.$$

Theorem 27.1.6 *Let μ and ν be probability measures on the Borel sets of $\mathbb{R}^p$ and suppose $\phi_\mu(t) = \phi_\nu(t)$. Then $\mu = \nu$.*

Proof: The proof is identical to the above. Just replace λ_X with μ and λ_Y with ν. ■

27.2 Conditional Probability

Here I will consider the concept of conditional probability depending on the theory of differentiation of general Radon measures. This leads to a different way of thinking about independence.

If X, Y are random vectors defined on a probability space having values in $\mathbb{R}^{p_1}$ and $\mathbb{R}^{p_2}$ respectively, and if E is a Borel set in the appropriate space, then (X, Y) is a random vector with values in $\mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$ and $\lambda_{(X,Y)}(E \times \mathbb{R}^{p_2}) = \lambda_X(E)$, $\lambda_{(X,Y)}(\mathbb{R}^{p_1} \times E) = \lambda_Y(E)$. Thus, by Theorem 19.8.3 on Page 538, there exist probability measures, denoted here by $\lambda_{X|y}$ and $\lambda_{Y|x}$, such that whenever E is a Borel set in $\mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$,

$$\int_{\mathbb{R}^{p_1}\times\mathbb{R}^{p_2}} \mathcal{X}_E d\lambda_{(X,Y)} = \int_{\mathbb{R}^{p_1}} \int_{\mathbb{R}^{p_2}} \mathcal{X}_E d\lambda_{Y|x} d\lambda_X,$$

and

$$\int_{\mathbb{R}^{p_1}\times\mathbb{R}^{p_2}} \mathcal{X}_E d\lambda_{(X,Y)} = \int_{\mathbb{R}^{p_2}} \int_{\mathbb{R}^{p_1}} \mathcal{X}_E d\lambda_{X|y} d\lambda_Y.$$

Definition 27.2.1 *Let X and Y be two random vectors defined on a probability space. The conditional probability measure of Y given X is the measure $\lambda_{Y|x}$ in the above. Similarly the conditional probability measure of X given Y is the measure $\lambda_{X|y}$.*

More generally, one can use the theory of slicing measures to consider any finite list of random vectors, $\{X_i\}$, defined on a probability space with $X_i \in \mathbb{R}^{p_i}$, and write the following for E a Borel set in $\prod_{i=1}^n \mathbb{R}^{p_i}$.

$$\int_{\mathbb{R}^{p_1}\times\cdots\times\mathbb{R}^{p_n}} \mathcal{X}_E d\lambda_{(X_1,\cdots,X_n)} = \int_{\mathbb{R}^{p_1}\times\cdots\times\mathbb{R}^{p_{n-1}}} \int_{\mathbb{R}^{p_n}} \mathcal{X}_E d\lambda_{X_n|(x_1,\cdots,x_{n-1})} d\lambda_{(X_1,\cdots,X_{n-1})}$$

$$= \int_{\mathbb{R}^{p_1}\times\cdots\times\mathbb{R}^{p_{n-2}}} \int_{\mathbb{R}^{p_{n-1}}} \int_{\mathbb{R}^{p_n}} \mathcal{X}_E d\lambda_{X_n|(x_1,\cdots,x_{n-1})} d\lambda_{X_{n-1}|(x_1,\cdots,x_{n-2})} d\lambda_{(X_1,\cdots,X_{n-2})}$$

$$\vdots$$

$$\int_{\mathbb{R}^{p_1}} \cdots \int_{\mathbb{R}^{p_n}} \mathcal{X}_E d\lambda_{X_n|(x_1,\cdots,x_{n-1})} d\lambda_{X_{n-1}|(x_1,\cdots,x_{n-2})} \cdots d\lambda_{X_2|x_1} d\lambda_{X_1}. \tag{27.1}$$

Obviously, this could have been done in any order in the iterated integrals by simply modifying the "given" variables, those occurring after the symbol $|$, to be those which have been integrated in an outer level of the iterated integral. For simplicity, write

$$\lambda_{X_n|(x_1,\cdots,x_{n-1})} = \lambda_{X_n|x_1,\cdots,x_{n-1}}$$

Definition 27.2.2 *Let $\{\boldsymbol{X}_1,\cdots,\boldsymbol{X}_n\}$ be random vectors defined on a probability space having values in $\mathbb{R}^{p_1},\cdots,\mathbb{R}^{p_n}$ respectively. The random vectors are independent if for every E a Borel set in $\mathbb{R}^{p_1}\times\cdots\times\mathbb{R}^{p_n}$,*

$$\int_{\mathbb{R}^{p_1}\times\cdots\times\mathbb{R}^{p_n}} \mathcal{X}_E d\lambda_{(\boldsymbol{X}_1,\cdots,\boldsymbol{X}_n)}$$

$$= \int_{\mathbb{R}^{p_1}}\cdots\int_{\mathbb{R}^{p_n}} \mathcal{X}_E d\lambda_{\boldsymbol{X}_n} d\lambda_{\boldsymbol{X}_{n-1}}\cdots d\lambda_{\boldsymbol{X}_2} d\lambda_{\boldsymbol{X}_1} \tag{27.2}$$

and the iterated integration may be taken in any order. If $\mathscr{A}$ is any set of random vectors defined on a probability space, $\mathscr{A}$ is independent if any finite set of random vectors from $\mathscr{A}$ is independent.

Thus, the random vectors are independent exactly when the dependence on the givens in 27.1 can be dropped.

Does this amount to the same thing as discussed earlier? Suppose you have three random variables $\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}$. Let $A=\boldsymbol{X}^{-1}(E)$, $B=\boldsymbol{Y}^{-1}(F)$, $C=\boldsymbol{Z}^{-1}(G)$ where E,F,G are Borel sets. Thus these inverse images are typical sets in

$$\sigma(\boldsymbol{X}),\sigma(\boldsymbol{Y}),\sigma(\boldsymbol{Z})$$

respectively. First suppose that the random variables are independent in the earlier sense. Then

$$P(A\cap B\cap C)=P(A)P(B)P(C)$$

$$\begin{aligned} &= \int_{\mathbb{R}^{p_1}} \mathcal{X}_E(\boldsymbol{x})\,d\lambda_{\boldsymbol{X}} \int_{\mathbb{R}^{p_2}} \mathcal{X}_F(\boldsymbol{y})\,d\lambda_{\boldsymbol{Y}} \int_{\mathbb{R}^{p_3}} \mathcal{X}_G(\boldsymbol{z})\,d\lambda_{\boldsymbol{Z}} \\ &= \int_{\mathbb{R}^{p_1}}\int_{\mathbb{R}^{p_2}}\int_{\mathbb{R}^{p_3}} \mathcal{X}_E(\boldsymbol{x})\,\mathcal{X}_F(\boldsymbol{y})\,\mathcal{X}_G(\boldsymbol{z})\,d\lambda_{\boldsymbol{Z}} d\lambda_{\boldsymbol{Y}} d\lambda_{\boldsymbol{X}} \end{aligned}$$

Also

$$P(A\cap B\cap C)=\int_{\mathbb{R}^{p_1}\times\mathbb{R}^{p_2}\times\mathbb{R}^{p_3}} \mathcal{X}_E(\boldsymbol{x})\,\mathcal{X}_F(\boldsymbol{y})\,\mathcal{X}_G(\boldsymbol{z})\,d\lambda_{(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z})}$$

$$= \int_{\mathbb{R}^{p_1}}\int_{\mathbb{R}^{p_2}}\int_{\mathbb{R}^{p_3}} \mathcal{X}_E(\boldsymbol{x})\,\mathcal{X}_F(\boldsymbol{y})\,\mathcal{X}_G(\boldsymbol{z})\,d\lambda_{\boldsymbol{Z}|\boldsymbol{x}\boldsymbol{y}} d\lambda_{\boldsymbol{Y}|\boldsymbol{x}} d\lambda_{\boldsymbol{X}}$$

Thus

$$\begin{aligned} &\int_{\mathbb{R}^{p_1}}\int_{\mathbb{R}^{p_2}}\int_{\mathbb{R}^{p_3}} \mathcal{X}_E(\boldsymbol{x})\,\mathcal{X}_F(\boldsymbol{y})\,\mathcal{X}_G(\boldsymbol{z})\,d\lambda_{\boldsymbol{Z}} d\lambda_{\boldsymbol{Y}} d\lambda_{\boldsymbol{X}} \\ = &\int_{\mathbb{R}^{p_1}}\int_{\mathbb{R}^{p_2}}\int_{\mathbb{R}^{p_3}} \mathcal{X}_E(\boldsymbol{x})\,\mathcal{X}_F(\boldsymbol{y})\,\mathcal{X}_G(\boldsymbol{z})\,d\lambda_{\boldsymbol{Z}|\boldsymbol{x}\boldsymbol{y}} d\lambda_{\boldsymbol{Y}|\boldsymbol{x}} d\lambda_{\boldsymbol{X}} \end{aligned}$$

Now letting $G=\mathbb{R}^{p_3}$, it follows that

$$\int_{\mathbb{R}^{p_1}}\int_{\mathbb{R}^{p_2}} \mathcal{X}_E(\boldsymbol{x})\,\mathcal{X}_F(\boldsymbol{y})\,d\lambda_{\boldsymbol{Y}} d\lambda_{\boldsymbol{X}} = \int_{\mathbb{R}^{p_1}}\int_{\mathbb{R}^{p_2}} \mathcal{X}_E(\boldsymbol{x})\,\mathcal{X}_F(\boldsymbol{y})\,d\lambda_{\boldsymbol{Y}|\boldsymbol{x}} d\lambda_{\boldsymbol{X}}$$

By uniqueness of the slicing measures or an application of the Besikovitch differentiation theorem, it follows that for $\lambda_{\boldsymbol{X}}$ a.e. $\boldsymbol{x}$,

$$\lambda_{\boldsymbol{Y}}=\lambda_{\boldsymbol{Y}|\boldsymbol{x}}$$

Thus, using this in the above,

$$\begin{aligned} & \int_{\mathbb{R}^{p_1}}\int_{\mathbb{R}^{p_2}}\int_{\mathbb{R}^{p_3}} \mathcal{X}_E(\boldsymbol{x})\,\mathcal{X}_F(\boldsymbol{y})\,\mathcal{X}_G(\boldsymbol{z})\,d\lambda_{\boldsymbol{Z}}d\lambda_{\boldsymbol{Y}}d\lambda_{\boldsymbol{X}} \\ = & \int_{\mathbb{R}^{p_1}}\int_{\mathbb{R}^{p_2}}\int_{\mathbb{R}^{p_3}} \mathcal{X}_E(\boldsymbol{x})\,\mathcal{X}_F(\boldsymbol{y})\,\mathcal{X}_G(\boldsymbol{z})\,d\lambda_{\boldsymbol{Z}|\boldsymbol{xy}}d\lambda_{\boldsymbol{Y}}d\lambda_{\boldsymbol{X}} \end{aligned}$$

and also it reduces to

$$\begin{aligned} & \int_{\mathbb{R}^{p_1}\times\mathbb{R}^{p_2}}\int_{\mathbb{R}^{p_3}} \mathcal{X}_E(\boldsymbol{x})\,\mathcal{X}_F(\boldsymbol{y})\,\mathcal{X}_G(\boldsymbol{z})\,d\lambda_{\boldsymbol{Z}}d\lambda_{(\boldsymbol{X},\boldsymbol{Y})} \\ = & \int_{\mathbb{R}^{p_1}\times\mathbb{R}^{p_2}}\int_{\mathbb{R}^{p_3}} \mathcal{X}_E(\boldsymbol{x})\,\mathcal{X}_F(\boldsymbol{y})\,\mathcal{X}_G(\boldsymbol{z})\,d\lambda_{\boldsymbol{Z}|\boldsymbol{xy}}d\lambda_{(\boldsymbol{X},\boldsymbol{Y})} \end{aligned}$$

Now by uniqueness of the slicing measures again, for $\lambda_{(\boldsymbol{X},\boldsymbol{Y})}$ a.e. $(\boldsymbol{x},\boldsymbol{y})$, it follows that

$$\lambda_{\boldsymbol{Z}} = \lambda_{\boldsymbol{Z}|\boldsymbol{xy}}$$

Similar conclusions hold for $\lambda_{\boldsymbol{X}}, \lambda_{\boldsymbol{Y}}$. In each case, off a set of measure zero the distribution measures equal the slicing measures.

Conversely, if the distribution measures equal the slicing measures off sets of measure zero as described above, then it is obvious that the random variables are independent. The same reasoning applies for any number of random variables.

Thus this gives a different and more analytical way to think of independence of finitely many random variables. Clearly, the argument given above will apply to any finite set of random variables.

Proposition 27.2.3 *Equations 27.2 and 27.1 hold with $\mathcal{X}_E$ replaced by any nonnegative Borel measurable function and for any bounded continuous function or for any function in L^1.*

Proof: The two equations hold for simple functions in place of $\mathcal{X}_E$ and so an application of the monotone convergence theorem applied to an increasing sequence of simple functions converging pointwise to a given nonnegative Borel measurable function yields the conclusion of the proposition in the case of the nonnegative Borel function. For a bounded continuous function or one in L^1, one can apply the result just established to the positive and negative parts of the real and imaginary parts of the function.

Lemma 27.2.4 *Let $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n$ be random vectors with values in $\mathbb{R}^{p_1}, \cdots, \mathbb{R}^{p_n}$ respectively and let*

$$\boldsymbol{g} : \mathbb{R}^{p_1} \times \cdots \times \mathbb{R}^{p_n} \to \mathbb{R}^k$$

be Borel measurable. Then $\boldsymbol{g}(\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)$ is a random vector with values in $\mathbb{R}^k$ and if $h : \mathbb{R}^k \to [0,\infty)$, then

$$\int_{\mathbb{R}^k} h(\boldsymbol{y})\,d\lambda_{\boldsymbol{g}(\boldsymbol{X}_1,\cdots,\boldsymbol{X}_n)}(y) =$$

$$\int_{\mathbb{R}^{p_1}\times\cdots\times\mathbb{R}^{p_n}} h(\boldsymbol{g}(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n))\,d\lambda_{(\boldsymbol{X}_1,\cdots,\boldsymbol{X}_n)}. \tag{27.3}$$

If $\boldsymbol{X}_i$ is a random vector with values in $\mathbb{R}^{p_i}, i = 1,2,\cdots$ and if $\boldsymbol{g}_i : \mathbb{R}^{p_i} \to \mathbb{R}^{k_i}$, where $\boldsymbol{g}_i$ is Borel measurable, then the random vectors $\boldsymbol{g}_i(\boldsymbol{X}_i)$ are also independent whenever the $\boldsymbol{X}_i$ are independent.

Proof: First let E be a Borel set in $\mathbb{R}^k$. From the definition,

$$\begin{aligned}\lambda_{\boldsymbol{g}(\boldsymbol{X}_1,\cdots,\boldsymbol{X}_n)}(E) &= P(\boldsymbol{g}(\boldsymbol{X}_1,\cdots,\boldsymbol{X}_n)\in E)\\ &= P\left((\boldsymbol{X}_1,\cdots,\boldsymbol{X}_n)\in \boldsymbol{g}^{-1}(E)\right)=\lambda_{(\boldsymbol{X}_1,\cdots,\boldsymbol{X}_n)}\left(\boldsymbol{g}^{-1}(E)\right)\end{aligned}$$

$$\begin{aligned}\int_{\mathbb{R}^k}\mathscr{X}_E d\lambda_{\boldsymbol{g}(\boldsymbol{X}_1,\cdots,\boldsymbol{X}_n)} &= \int_{\mathbb{R}^{p_1}\times\cdots\times\mathbb{R}^{p_n}}\mathscr{X}_{\boldsymbol{g}^{-1}(E)}d\lambda_{(\boldsymbol{X}_1,\cdots,\boldsymbol{X}_n)}\\ &= \int_{\mathbb{R}^{p_1}\times\cdots\times\mathbb{R}^{p_n}}\mathscr{X}_E(\boldsymbol{g}(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n))\,d\lambda_{(\boldsymbol{X}_1,\cdots,\boldsymbol{X}_n)}.\end{aligned}$$

This proves 27.3 in the case when h is $\mathscr{X}_E$. To prove it in the general case, approximate the nonnegative Borel measurable function with simple functions for which the formula is true, and use the monotone convergence theorem.

It remains to prove the last assertion that functions of independent random vectors are also independent random vectors. Let E be a Borel set in $\mathbb{R}^{k_1}\times\cdots\times\mathbb{R}^{k_n}$. Then for

$$\pi_i(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n)\equiv\boldsymbol{x}_i,$$

$$\begin{aligned}&\int_{\mathbb{R}^{k_1}\times\cdots\times\mathbb{R}^{k_n}}\mathscr{X}_E d\lambda_{(\boldsymbol{g}_1(\boldsymbol{X}_1),\cdots,\boldsymbol{g}_n(\boldsymbol{X}_n))}\\ &\equiv\int_{\mathbb{R}^{p_1}\times\cdots\times\mathbb{R}^{p_n}}\mathscr{X}_E\circ(\boldsymbol{g}_1\circ\pi_1,\cdots,\boldsymbol{g}_n\circ\pi_n)\,d\lambda_{(\boldsymbol{X}_1,\cdots,\boldsymbol{X}_n)}\\ &= \int_{\mathbb{R}^{p_1}}\cdots\int_{\mathbb{R}^{p_n}}\mathscr{X}_E\circ(\boldsymbol{g}_1\circ\pi_1,\cdots,\boldsymbol{g}_n\circ\pi_n)\,d\lambda_{\boldsymbol{X}_n}\cdots d\lambda_{\boldsymbol{X}_1}\\ &= \int_{\mathbb{R}^{k_1}}\cdots\int_{\mathbb{R}^{k_n}}\mathscr{X}_E d\lambda_{\boldsymbol{g}_n(\boldsymbol{X}_n)}\cdots d\lambda_{\boldsymbol{g}_1(\boldsymbol{X}_1)}\ \blacksquare\end{aligned}$$

Of course if $X_i, i=1,2,...,n$ are independent, this means the σ algebras $\sigma(X_i)$ are independent. Now $\sigma(g_i\circ X_i)\subseteq\sigma(X_i)$ because

$$(g_i\circ X_i)^{-1}(\text{Borel set})=X_i^{-1}\left(g_i^{-1}(\text{Borel set})\right)=X_i^{-1}(\text{Borel set})\in\sigma(X_i)$$

and so the variables $g_i\circ X_i$, $i=1,2,...,n$ are independent. I think this is a more direct way of seeing this second claim.

Proposition 27.2.5 *Let $\nu_1,\cdots,\nu_n$ be Radon probability measures defined on $\mathbb{R}^p$. Then there exists a probability space and independent random vectors*

$$\{\boldsymbol{X}_1,\cdots,\boldsymbol{X}_n\}$$

defined on this probability space such that $\lambda_{\boldsymbol{X}_i}=\nu_i$.

Proof: Let $(\Omega,\mathscr{S},P)\equiv((\mathbb{R}^p)^n,\mathscr{S}_1\times\cdots\times\mathscr{S}_n,\nu_1\times\cdots\times\nu_n)$ where this is just the product σ algebra and product measure which satisfies the following for measurable rectangles.

$$(\nu_1\times\cdots\times\nu_n)\left(\prod_{i=1}^n E_i\right)=\prod_{i=1}^n\nu_i(E_i).$$

Now let $\boldsymbol{X}_i(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_i,\cdots,\boldsymbol{x}_n)=\boldsymbol{x}_i$.

Then from the definition, if E is a Borel set in $\mathbb{R}^p$,

$$\lambda_{\boldsymbol{X}_i}(E) \equiv P\{\boldsymbol{X}_i \in E\}$$

$$= (\nu_1 \times \cdots \times \nu_n)(\mathbb{R}^p \times \cdots \times E \times \cdots \times \mathbb{R}^p) = \nu_i(E).$$

This defines the random vectors $\{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n\}$ such that $\lambda_{\boldsymbol{X}_i} = \nu_i$ on all Borel sets. Are these random vectors independent? Let $\mathscr{M}$ consist of all Borel sets of $(\mathbb{R}^p)^n$ such that

$$\int_{\mathbb{R}^p} \cdots \int_{\mathbb{R}^p} \mathscr{X}_E(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)\, d\lambda_{\boldsymbol{X}_1} \cdots d\lambda_{\boldsymbol{X}_n} = \int_{(\mathbb{R}^p)^n} \mathscr{X}_E d\lambda_{(\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)}.$$

It is clear that $\mathscr{M}$ contains all products of Borel sets $\prod_{i=1}^n E_i$ which is a π system called $\mathscr{K}$. It is also clearly closed with respect to countable disjoint unions and complements. Thus. by the lemma on π systems, Lemma 9.3.2, it contains $\sigma(\mathscr{K})$ which is the Borel sets because it contains all open sets. Therefore, the given random vectors are independent because you can dispense with the givens. ■

The following Lemma was proved earlier in a different way.

Lemma 27.2.6 *If $\{X_i\}_{i=1}^n$ are independent random variables having values in $\mathbb{R}$,*

$$E\left(\prod_{i=1}^n X_i\right) = \prod_{i=1}^n E(X_i).$$

Proof: By Lemma 27.2.4 and denoting by P the product, $\prod_{i=1}^n X_i$,

$$\begin{aligned} E\left(\prod_{i=1}^n X_i\right) &= \int_{\mathbb{R}} z\, d\lambda_P(z) = \int_{\mathbb{R}^n} \prod_{i=1}^n x_i\, d\lambda_{(X_1, \cdots, X_n)} \\ &= \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \prod_{i=1}^n x_i\, d\lambda_{X_1} \cdots d\lambda_{X_n} = \prod_{i=1}^n E(X_i). \blacksquare \end{aligned}$$

27.3 Conditional Expectation, Sub-martingales

This concept is developed more later when conditional expectation relative to a σ algebra is discussed. However, I think there tends to be a disconnect between that more abstract idea and what we usually think of where conditional expectation involves expectation in which conditional probability is used, where a random variable is "given" to have a particular value, as described above. Certainly this is the case in beginning treatments of probability as an application of combinatorics or in beginning statistics classes. Therefore, to tie this in to this more elementary way of thinking, I will present conditional expectation in terms of conditional probability as defined above, where values of random variables are given. This leads to a rudimentary treatment of the sub-martingale convergence theorem presented here. However, the more abstract version presented later is much less difficult I think, because it is dependent on the Radon Nikodym theorem rather than the very difficult Besicovitch differentiation theory of Radon measures.

Definition 27.3.1 *Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be random vectors having values in $\mathbb{F}^{p_1}$ and $\mathbb{F}^{p_2}$ respectively. Then if $\int |\boldsymbol{x}|\, d\lambda_{\boldsymbol{X}|\boldsymbol{y}}(x) < \infty$, define $E(\boldsymbol{X}|\boldsymbol{y}) \equiv \int \boldsymbol{x}\, d\lambda_{\boldsymbol{X}|\boldsymbol{y}}(x)$.*

Proposition 27.3.2 *Suppose* $\int_{\mathbb{F}^{p_1}\times\mathbb{F}^{p_2}} |\boldsymbol{x}|\, d\lambda_{(\boldsymbol{X},\boldsymbol{Y})}(x) < \infty$. *Then* $E(\boldsymbol{X}|\boldsymbol{y})$ *exists for* $\lambda_{\boldsymbol{Y}}$ *a.e.* $\boldsymbol{y}$ *and*

$$\int_{\mathbb{F}^{p_2}} E(\boldsymbol{X}|\boldsymbol{y})\, d\lambda_{\boldsymbol{Y}} = \int_{\mathbb{F}^{p_1}} \boldsymbol{x}\, d\lambda_{\boldsymbol{X}}(x) = E(\boldsymbol{X}).$$

Proof: $\infty > \int_{\mathbb{F}^{p_1}\times\mathbb{F}^{p_2}} |\boldsymbol{x}|\, d\lambda_{(\boldsymbol{X},\boldsymbol{Y})} = \int_{\mathbb{F}^{p_2}} \int_{\mathbb{F}^{p_1}} |\boldsymbol{x}|\, d\lambda_{\boldsymbol{X}|\boldsymbol{y}}(x)\, d\lambda_{\boldsymbol{Y}}(y)$ and so

$$\int_{\mathbb{F}^{p_1}} |\boldsymbol{x}|\, d\lambda_{\boldsymbol{X}|\boldsymbol{y}}(x) < \infty,$$

$\lambda_{\boldsymbol{Y}}$ a.e. Now $\int_{\mathbb{F}^{p_2}} E(\boldsymbol{X}|\boldsymbol{y})\, d\lambda_{\boldsymbol{Y}} =$

$$\begin{aligned} &= \int_{\mathbb{F}^{p_2}} \int_{\mathbb{F}^{p_1}} \boldsymbol{x}\, d\lambda_{\boldsymbol{X}|\boldsymbol{y}}(x)\, d\lambda_{\boldsymbol{Y}}(y) = \int_{\mathbb{F}^{p_1}\times\mathbb{F}^{p_2}} \boldsymbol{x}\, d\lambda_{(\boldsymbol{X},\boldsymbol{Y})} \\ &= \int_{\mathbb{F}^{p_1}} \int_{\mathbb{F}^{p_2}} \boldsymbol{x}\, d\lambda_{\boldsymbol{Y}|\boldsymbol{x}}(y)\, d\lambda_{\boldsymbol{X}}(x) = \int_{\mathbb{F}^{p_2}} \boldsymbol{x}\, d\lambda_{\boldsymbol{X}}(x) = E(\boldsymbol{X}) \blacksquare \end{aligned}$$

Definition 27.3.3 *Let* $\{X_n\}$ *be any sequence, finite or infinite, of random variables with values in* $\mathbb{R}$ *which are defined on some probability space,* $(\Omega, \mathscr{S}, P)$. *We say* $\{X_n\}$ *is a martingale if*

$$E(X_n|x_{n-1},\cdots,x_1) = x_{n-1}$$

and we say $\{X_n\}$ *is a sub-martingale if*

$$E(X_n|x_{n-1},\cdots,x_1) \geq x_{n-1}.$$

Here is a very useful lemma, the last statement known as Jensen's inequality. It might help in understanding the first part of this to draw a picture of a convex function and think in terms of the various slopes of secant lines.

Lemma 27.3.4 *If* $\phi : \mathbb{R} \to \mathbb{R}$ *is convex, then* ϕ *is continuous. Also, if* ϕ *is convex,* $\mu(\Omega) = 1$, *and* $f, \phi(f) : \Omega \to \mathbb{R}$ *are in* $L^1(\Omega)$, *then* $\phi(\int_\Omega f\, du) \leq \int_\Omega \phi(f) d\mu$.

Proof: Let $x < y < z$. Then $y = x + \mu(z-x) = (1-\mu)x + \mu z$ for some $\mu \in (0,1)$. Then

$$\begin{aligned} \frac{\phi(y)-\phi(x)}{y-x} &= \frac{\phi((1-\mu)x+\mu z)-\phi(x)}{y-x} \\ &\leq \frac{(1-\mu)\phi(x)+\mu\phi(z)-\phi(x)}{((1-\mu)x+\mu z)-x} = \frac{\phi(z)-\phi(x)}{z-x} \end{aligned} \tag{27.4}$$

Then from this inequality,

$$\phi(y) \leq \phi(x) + \frac{\phi(z)-\phi(x)}{z-x}(y-x)$$

Now let $\lambda \equiv \inf\left\{\frac{\phi(y)-\phi(x)}{y-x} : y > x\right\}$. Then $\frac{\phi(y)-\phi(x)}{y-x} \geq \lambda$ and so $\phi(y) \geq \lambda(y-x) + \phi(x)$. Thus

$$\frac{\phi(z)-\phi(x)}{z-x}(y-x) \geq \phi(y)-\phi(x) \geq \lambda(y-x) \tag{27.5}$$

and so ϕ is continuous from the right at x. But x was arbitrary so ϕ is continuous from the right.

$$\frac{\phi(z)-\phi(y)}{z-y} \geq \frac{\phi(z)-((1-\mu)\phi(x)+\mu\phi(z))}{z-((1-\mu)x+\mu z)}$$

$$= \frac{(1-\mu)(\phi(z)-\phi(x))}{(1-\mu)(z-x)} = \frac{(\phi(z)-\phi(x))}{z-x} \tag{27.6}$$

Using similar reasoning ϕ is continuous from the left at z for any z. Thus ϕ is a continuous function.

Now consider the other claim about Jensen's inequality. Letting λ be as described above, it was shown that $\phi(x) \leq \phi(y) + (-\lambda)(y-x)$ or in other words, for $y > x, \phi(x) \leq \phi(y) + \lambda(x-y)$. Does this inequality continue to hold if $y < x$? From 27.4 and 27.6, $\frac{\phi(z)-\phi(y)}{z-y} \geq \frac{\phi(y)-\phi(x)}{y-x}$ whenever $y \in (x,z)$. Letting $\hat{y} < x, y > x$ it follows that

$$\frac{\phi(\hat{y})-\phi(x)}{\hat{y}-x} \leq \frac{\phi(y)-\phi(x)}{y-x}$$

and so $\frac{\phi(\hat{y})-\phi(x)}{\hat{y}-x} = \frac{\phi(x)-\phi(\hat{y})}{x-\hat{y}} \leq \lambda$ and so $\phi(x) \leq \phi(\hat{y}) + \lambda(x-\hat{y})$ and so, for all $y \in \mathbb{R}$, $\phi(x) \leq \phi(y) + \lambda(x-y)$. Now from this it follows that, letting $x = \int_\Omega f d\mu$, and using the λ whose existence was just established, for each ω,

$$\phi\left(\int_\Omega f d\mu\right) \leq \phi(f) + \lambda\left(\int_\Omega f d\mu - f(\omega)\right)$$

Do $\int_\Omega d\mu$ to both sides and use $\mu(\Omega) = 1$. Thus

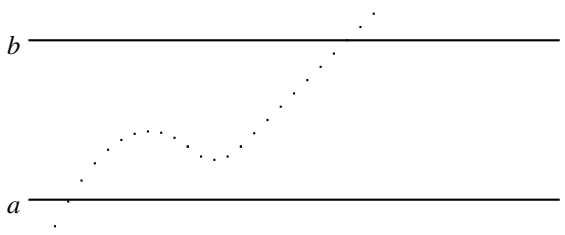
$$\phi\left(\int_\Omega f d\mu\right) \leq \int_\Omega \phi(f) d\mu + \lambda\left(\int_\Omega f d\mu - \int_\Omega f d\mu\right) = \int_\Omega \phi(f) d\mu.$$

There are no difficulties with measurability because ϕ is continuous. ■

Next is the notion of an upcrossing.

Definition 27.3.5 *Let $\{x_i\}_{i=1}^I$ be any sequence of real numbers, $I \leq \infty$. Define an increasing sequence of integers $\{m_k\}$ as follows. m_1 is the first integer ≥ 1 such that $x_{m_1} \leq a$, m_2 is the first integer larger than m_1 such that $x_{m_2} \geq b$, m_3 is the first integer larger than m_2 such that $x_{m_3} \leq a$, etc. Then each sequence, $\left\{x_{m_{2k-1}}, \cdots, x_{m_{2k}}\right\}$, is called an upcrossing of $[a,b]$.*

Here is a picture of an upcrossing.



Proposition 27.3.6 *Let $\{X_i\}_{i=1}^n$ be a finite sequence of real random variables defined on Ω where $(\Omega, \mathcal{S}, P)$ is a probability space. Let $U_{[a,b]}(\omega)$ denote the number of upcrossings of $X_i(\omega)$ of the interval $[a,b]$. Then $U_{[a,b]}$ is a random variable, in other words, a nonnegative measurable function.*

Proof: Let $X_0(\omega) \equiv a+1$, let $Y_0(\omega) \equiv 0$, and let $Y_k(\omega)$ remain 0 for $k = 0, \cdots, l$ until $X_l(\omega) \leq a$. When this happens (if ever), $Y_{l+1}(\omega) \equiv 1$. Then let $Y_i(\omega)$ remain 1 for $i = l+1, \cdots, r$ until $X_r(\omega) \geq b$ when $Y_{r+1}(\omega) \equiv 0$. Let $Y_k(\omega)$ remain 0 for $k \geq r+1$ until $X_k(\omega) \leq a$ when $Y_k(\omega) \equiv 1$ and continue in this way. Thus the upcrossings of $X_i(\omega)$ are identified as unbroken strings of ones with a zero at each end, with the possible exception of the last string of ones which may be missing the zero at the upper end and may or may not be an upcrossing.

Note also that Y_0 is measurable because it is identically equal to 0 and that if Y_k is measurable, then Y_{k+1} is measurable because the only change in going from k to $k+1$ is a change from 0 to 1 or from 1 to 0 on a measurable set determined by X_k. Now let

$$Z_k(\omega) = \begin{cases} 1 \text{ if } Y_k(\omega) = 1 \text{ and } Y_{k+1}(\omega) = 0, \\ 0 \text{ otherwise,} \end{cases}$$

if $k < n$ and

$$Z_n(\omega) = \begin{cases} 1 \text{ if } Y_n(\omega) = 1 \text{ and } X_n(\omega) \geq b, \\ 0 \text{ otherwise.} \end{cases}$$

Thus $Z_k(\omega) = 1$ exactly when an upcrossing has been completed and each Z_i is a random variable.

$$U_{[a,b]}(\omega) = \sum_{k=1}^{n} Z_k(\omega)$$

so $U_{[a,b]}$ is a random variable as claimed. ■

The following corollary collects some key observations found in the above construction.

Corollary 27.3.7 *$U_{[a,b]}(\omega) \leq$ the number of unbroken strings of ones in the sequence $\{Y_k(\omega)\}$ there being at most one unbroken string of ones which produces no upcrossing. Also*

$$Y_i(\omega) = \psi_i\left(\left\{X_j(\omega)\right\}_{j=1}^{i-1}\right), \tag{27.7}$$

where ψ_i is some function of the past values of $X_j(\omega)$.

Lemma 27.3.8 *(upcrossing lemma) Let $\{X_i\}_{i=1}^{n}$ be a sub-martingale and suppose*

$$E(|X_n|) < \infty.$$

Then

$$E\left(U_{[a,b]}\right) \leq \frac{E(|X_n|) + |a|}{b-a}.$$

Proof: Let $\phi(x) \equiv a + (x-a)^+$. Thus ϕ is a convex and increasing function.

$$\phi(X_{k+r}) - \phi(X_k) = \sum_{i=k+1}^{k+r} \phi(X_i) - \phi(X_{i-1})$$

$$= \sum_{i=k+1}^{k+r} (\phi(X_i) - \phi(X_{i-1}))\, Y_i + \sum_{i=k+1}^{k+r} (\phi(X_i) - \phi(X_{i-1}))\,(1 - Y_i).$$

The upcrossings of $\phi(X_i)$ are exactly the same as the upcrossings of X_i and from 27.7,

$$E\left(\sum_{i=k+1}^{k+r} (\phi(X_i) - \phi(X_{i-1}))\,(1 - Y_i)\right)$$

$$= \sum_{i=k+1}^{k+r} \int_{\mathbb{R}^i} (\phi(x_i) - \phi(x_{i-1})) \left(1 - \psi_i\left(\{x_j\}_{j=1}^{i-1}\right)\right) d\lambda_{(X_1,\cdots,X_i)}$$

$$= \sum_{i=k+1}^{k+r} \int_{\mathbb{R}^{i-1}} \int_{\mathbb{R}} (\phi(x_i) - \phi(x_{i-1}))\cdot$$

$$\left(1 - \psi_i\left(\{x_j\}_{j=1}^{i-1}\right)\right) d\lambda_{X_i|x_1\cdots x_{i-1}} d\lambda_{(X_1,\cdots,X_{i-1})}$$

$$= \sum_{i=k+1}^{k+r} \int_{\mathbb{R}^{i-1}} \left(1 - \psi_i\left(\{x_j\}_{j=1}^{i-1}\right)\right)\cdot$$

$$\int_{\mathbb{R}} (\phi(x_i) - \phi(x_{i-1})) d\lambda_{X_i|x_1\cdots x_{i-1}} d\lambda_{(X_1,\cdots,X_{i-1})} \tag{27.8}$$

By Jensen's inequality, Lemma 27.3.4 and that this is a sub-martingale, and that ϕ is increasing in addition to being convex,

$$\int_{\mathbb{R}} \phi(x_i) d\lambda_{X_i|x_1\cdots x_{i-1}} \geq \phi\left(\int_{\mathbb{R}} x_i d\lambda_{X_i|x_1\cdots x_{i-1}}\right) \geq \phi(x_{i-1}).$$

Therefore, from 27.8, $E\left(\sum_{i=k+1}^{k+r} (\phi(X_i) - \phi(X_{i-1}))(1 - Y_i)\right) \geq$

$$\sum_{i=k+1}^{k+r} \int_{\mathbb{R}^{i-1}} \left(1 - \psi_i\left(\{x_j\}_{j=1}^{i-1}\right)\right) (\phi(x_{i-1}) - \phi(x_{i-1})) d\lambda_{(X_1,\cdots,X_{i-1})} = 0$$

Now let the unbroken strings of ones for $\{Y_i(\omega)\}$ be

$$\{k_1,\cdots,k_1+r_1\},\{k_2,\cdots,k_2+r_2\},\cdots,\{k_m,\cdots,k_m+r_m\} \tag{27.9}$$

where $m = V(\omega) \equiv$ the number of unbroken strings of ones in the sequence $\{Y_i(\omega)\}$. By Corollary 27.3.7 $V(\omega) \geq U_{[a,b]}(\omega)$.

$$\phi(X_n(\omega)) - \phi(X_1(\omega))$$

$$= \sum_{k=1}^{n} (\phi(X_k(\omega)) - \phi(X_{k-1}(\omega))) Y_k(\omega)$$

$$+ \sum_{k=1}^{n} (\phi(X_k(\omega)) - \phi(X_{k-1}(\omega))) (1 - Y_k(\omega)).$$

Summing the first sum over the unbroken strings of ones (the terms in which $Y_i(\omega) = 0$ contribute nothing), and observing that for $x > a, \phi(x) = x$,

$$\phi(X_n(\omega)) - \phi(X_1(\omega)) \geq U_{[a,b]}(\omega)(b-a) + 0 +$$

$$\sum_{k=1}^{n} (\phi(X_k(\omega)) - \phi(X_{k-1}(\omega))) (1 - Y_k(\omega)) \tag{27.10}$$

where the zero on the right side results from a string of ones which does not produce an upcrossing. It is here that we use $\phi(x) \geq a$. Such a string begins with $\phi(X_k(\omega)) = a$ and results in an expression of the form $\phi(X_{k+m}(\omega)) - \phi(X_k(\omega)) \geq 0$ since $\phi(X_{k+m}(\omega)) \geq a$.

If we had not replaced X_k with $\phi(X_k)$, it would have been possible for $\phi(X_{k+m}(\omega))$ to be less than a and the zero in the above could have been a negative number.

Therefore from 27.10,

$$\begin{aligned}(b-a)E\left(U_{[a,b]}\right) &\leq E(\phi(X_n)-\phi(X_1)) \leq E(\phi(X_n)-a)\\ &= E\left((X_n-a)^+\right) \leq |a|+E(|X_n|) \blacksquare\end{aligned}$$

With this estimate, the amazing sub-martingale convergence theorem follows. This incredible theorem says that a bounded in L^1 sub-martingale must converge a.e.

Theorem 27.3.9 *(sub-martingale convergence theorem) Let $\{X_i\}_{i=1}^{\infty}$ be a sub-martingale with $K \equiv \sup\{E(|X_n|) : n \geq 1\} < \infty$. Then there exists a random variable X_∞, such that $E(|X_\infty|) \leq K$ and $\lim_{n\to\infty} X_n(\omega) = X_\infty(\omega)$ a.e.*

Proof: Let $a,b \in \mathbb{Q}$ and let $a < b$. Let $U_{[a,b]}^n(\omega)$ be the number of upcrossings of $\{X_i(\omega)\}_{i=1}^n$. Then let

$$U_{[a,b]}(\omega) \equiv \lim_{n\to\infty} U_{[a,b]}^n(\omega) = \text{ number of upcrossings of } \{X_i\}.$$

By the upcrossing lemma, $E\left(U_{[a,b]}^n\right) \leq \frac{E(|X_n|)+|a|}{b-a} \leq \frac{K+|a|}{b-a}$ and so by the monotone convergence theorem, $E\left(U_{[a,b]}\right) \leq \frac{K+|a|}{b-a} < \infty$ which shows $U_{[a,b]}(\omega)$ is finite a.e., for all $\omega \notin S_{[a,b]}$ where $P\left(S_{[a,b]}\right) = 0$. Define $S \equiv \cup\left\{S_{[a,b]} : a,b \in \mathbb{Q}, a < b\right\}$. Then $P(S) = 0$ and if $\omega \notin S$, $\{X_k\}_{k=1}^{\infty}$ has only finitely many upcrossings of every interval having rational endpoints. Thus, for $\omega \notin S$,

$$\limsup_{k\to\infty} X_k(\omega) = \liminf_{k\to\infty} X_k(\omega) = \lim_{k\to\infty} X_k(\omega) \equiv X_\infty(\omega).$$

Letting $X_\infty(\omega) = 0$ for $\omega \in S$, Fatou's lemma implies

$$\int_\Omega |X_\infty|\,dP = \int_\Omega \liminf_{n\to\infty} |X_n|\,dP \leq \liminf_{n\to\infty} \int_\Omega |X_n|\,dP \leq K \blacksquare$$

27.4 Characteristic Functions and Independence

There is a way to tell if random vectors are independent by using their characteristic functions.

Proposition 27.4.1 *If $\boldsymbol{X}_i$ is a random vector having values in $\mathbb{R}^{p_i}$, then the random vectors are independent if and only if*

$$E\left(e^{iP}\right) = \prod_{j=1}^{n} E\left(e^{i\boldsymbol{t}_j\cdot\boldsymbol{X}_j}\right)$$

where $P \equiv \sum_{j=1}^n \boldsymbol{t}_j \cdot \boldsymbol{X}_j$ for $\boldsymbol{t}_j \in \mathbb{R}^{p_j}$.

The proof of this proposition will depend on the following lemma.

Lemma 27.4.2 *Let $\boldsymbol{Y}$ be a random vector with values in $\mathbb{R}^p$ and let f be bounded and measurable with respect to the Radon measure $\lambda_{\boldsymbol{Y}}$, and satisfy*

$$\int f(\boldsymbol{y}) e^{i\boldsymbol{t}\cdot\boldsymbol{y}} d\lambda_{\boldsymbol{Y}} = 0$$

for all $\boldsymbol{t} \in \mathbb{R}^p$. Then $f(\boldsymbol{y}) = 0$ for $\lambda_{\boldsymbol{Y}}$ a.e. $\boldsymbol{y}$.

Proof: You could write the following for $\phi \in \mathscr{G}$

$$\int \phi(\boldsymbol{t}) \int f(\boldsymbol{y}) e^{i\boldsymbol{t}\cdot\boldsymbol{y}} d\lambda_{\boldsymbol{Y}} dt = 0 = \int f(\boldsymbol{y}) \left(\int \phi(\boldsymbol{t}) e^{i\boldsymbol{t}\cdot\boldsymbol{y}} dt \right) d\lambda_{\boldsymbol{Y}}$$

Recall that the inverse Fourier transform maps $\mathscr{G}$ onto $\mathscr{G}$. Hence $\int f(\boldsymbol{y}) \psi(\boldsymbol{y}) d\lambda_{\boldsymbol{Y}} = 0$ for all $\psi \in \mathscr{G}$. Thus this is also so for every $\psi \in C_0^\infty(\mathbb{R}^p) \supseteq C_c^\infty(\mathbb{R}^p)$ by an obvious application of the Stone Weierstrass theorem. Let $\{\phi_k\}$ be a sequence of functions in $C_c^\infty(\mathbb{R}^p)$ which converges to

$$\operatorname{sgn}(f) \equiv \begin{cases} \bar{f}/|f| \text{ if } f \neq 0 \\ 0 \text{ if } f = 0 \end{cases}$$

pointwise and in $L^1(\mathbb{R}^p, \lambda_{\boldsymbol{Y}})$, each $|\phi_k| \leq 2$. Then for any $\psi \in C_0^\infty(\mathbb{R}^p)$,

$$0 = \int f(\boldsymbol{y}) \phi_n(\boldsymbol{y}) \psi(\boldsymbol{y}) d\lambda_{\boldsymbol{Y}} \to \int |f(\boldsymbol{y})| \psi(\boldsymbol{y}) d\lambda_{\boldsymbol{Y}}$$

Also, the above holds for any $\psi \in C_c(\mathbb{R}^p)$ as can be seen by taking such a ψ and convolving with a mollifier. By the Riesz representation theorem, $f(\boldsymbol{y}) = 0$ $\lambda_{\boldsymbol{Y}}$ a.e. (The measure $\mu(E) \equiv \int_E |f(\boldsymbol{y})| d\lambda_{\boldsymbol{Y}}$ equals 0.) ■

Proof of the proposition: If the $\boldsymbol{X}_j$ are independent, the formula follows from Lemma 27.2.6 and Lemma 27.2.4.

Now suppose the formula holds. Thus $\prod_{j=1}^n E\left(e^{i\boldsymbol{t}_j\cdot\boldsymbol{X}_j}\right) =$

$$\int_{\mathbb{R}^{p_n}} \cdots \int_{\mathbb{R}^{p_2}} \int_{\mathbb{R}^{p_1}} e^{i\boldsymbol{t}_1\cdot\boldsymbol{x}_1} e^{i\boldsymbol{t}_2\cdot\boldsymbol{x}_2} \cdots e^{i\boldsymbol{t}_n\cdot\boldsymbol{x}_n} d\lambda_{\boldsymbol{X}_1} d\lambda_{\boldsymbol{X}_2} \cdots d\lambda_{\boldsymbol{X}_n} = E\left(e^{iP}\right)$$

$$= \int_{\mathbb{R}^{p_n}} \cdots \int_{\mathbb{R}^{p_2}} \int_{\mathbb{R}^{p_1}} e^{i\boldsymbol{t}_1\cdot\boldsymbol{x}_1} e^{i\boldsymbol{t}_2\cdot\boldsymbol{x}_2} \cdots e^{i\boldsymbol{t}_n\cdot\boldsymbol{x}_n} d\lambda_{\boldsymbol{X}_1|\boldsymbol{x}_2\cdots\boldsymbol{x}_n} d\lambda_{\boldsymbol{X}_2|\boldsymbol{x}_3\cdots\boldsymbol{x}_n} \cdots d\lambda_{\boldsymbol{X}_n}. \quad (27.11)$$

Then from the above Lemma 27.4.2, the following equals 0 for $\lambda_{\boldsymbol{X}_n}$ a.e. $\boldsymbol{x}_n$.

$$\int_{\mathbb{R}^{p_{n-1}}} \cdots \int_{\mathbb{R}^{p_2}} \int_{\mathbb{R}^{p_1}} e^{i\boldsymbol{t}_1\cdot\boldsymbol{x}_1} e^{i\boldsymbol{t}_2\cdot\boldsymbol{x}_2} \cdots e^{i\boldsymbol{t}_{n-1}\cdot\boldsymbol{x}_{n-1}} d\lambda_{\boldsymbol{X}_1} d\lambda_{\boldsymbol{X}_2} \cdots d\lambda_{\boldsymbol{X}_{n-1}} -$$

$$\int_{\mathbb{R}^{p_{n-1}}} \cdots \int_{\mathbb{R}^{p_2}} \int_{\mathbb{R}^{p_1}} e^{i\boldsymbol{t}_1\cdot\boldsymbol{x}_1} e^{i\boldsymbol{t}_2\cdot\boldsymbol{x}_2} \cdots e^{i\boldsymbol{t}_{n-1}\cdot\boldsymbol{x}_{n-1}} d\lambda_{\boldsymbol{X}_1|\boldsymbol{x}_2\cdots\boldsymbol{x}_n} d\lambda_{\boldsymbol{X}_2|\boldsymbol{x}_3\cdots\boldsymbol{x}_n} \cdots d\lambda_{\boldsymbol{X}_{n-1}|\boldsymbol{x}_n}$$

Let $\boldsymbol{t}_i = \boldsymbol{0}$ for $i = 1, 2, \cdots, n-2$. Then this implies

$$\int_{\mathbb{R}^{p_{n-1}}} e^{i\boldsymbol{t}_{n-1}\cdot\boldsymbol{x}_{n-1}} d\lambda_{\boldsymbol{X}_{n-1}} = \int_{\mathbb{R}^{p_{n-1}}} e^{i\boldsymbol{t}_{n-1}\cdot\boldsymbol{x}_{n-1}} d\lambda_{\boldsymbol{X}_{n-1}|\boldsymbol{x}_n}$$

By the fact that the characteristic function determines the distribution measure, Theorem 27.1.4, it follows that for these $\boldsymbol{x}_n$ off a set of $\lambda_{\boldsymbol{X}_n}$ measure zero, $\lambda_{\boldsymbol{X}_{n-1}} = \lambda_{\boldsymbol{X}_{n-1}|\boldsymbol{x}_n}$. Returning to 27.11, one can replace $\lambda_{\boldsymbol{X}_{n-1}|\boldsymbol{x}_n}$ with $\lambda_{\boldsymbol{X}_{n-1}}$ to obtain

$$\int_{\mathbb{R}^{p_n}} \cdots \int_{\mathbb{R}^{p_2}} \int_{\mathbb{R}^{p_1}} e^{i\boldsymbol{t}_1\cdot\boldsymbol{x}_1} e^{i\boldsymbol{t}_2\cdot\boldsymbol{x}_2} \cdots e^{i\boldsymbol{t}_n\cdot\boldsymbol{x}_n} d\lambda_{\boldsymbol{X}_1} d\lambda_{\boldsymbol{X}_2} \cdots d\lambda_{\boldsymbol{X}_{n-1}} d\lambda_{\boldsymbol{X}_n}$$

$$= \int_{\mathbb{R}^{p_n}} \cdots \int_{\mathbb{R}^{p_2}} \int_{\mathbb{R}^{p_1}} e^{i\boldsymbol{t}_1\cdot\boldsymbol{x}_1} e^{i\boldsymbol{t}_2\cdot\boldsymbol{x}_2} \cdots e^{i\boldsymbol{t}_n\cdot\boldsymbol{x}_n} \cdot d\lambda_{\boldsymbol{X}_1|\boldsymbol{x}_2\cdots\boldsymbol{x}_n} d\lambda_{\boldsymbol{X}_2|\boldsymbol{x}_3\cdots\boldsymbol{x}_n} \cdots d\lambda_{\boldsymbol{X}_{n-1}} d\lambda_{\boldsymbol{X}_n}$$

Next let $\boldsymbol{t}_n = 0$ and applying the above Lemma 27.4.2 again, this implies that for $\lambda_{\boldsymbol{X}_{n-1}}$ a.e. $\boldsymbol{x}_{n-1}$, the following equals 0.

$$\int_{\mathbb{R}^{p_{n-2}}} \cdots \int_{\mathbb{R}^{p_2}} \int_{\mathbb{R}^{p_1}} e^{i\boldsymbol{t}_1\cdot\boldsymbol{x}_1} e^{i\boldsymbol{t}_2\cdot\boldsymbol{x}_2} \cdots e^{i\boldsymbol{t}_{n-2}\cdot\boldsymbol{x}_{n-2}} d\lambda_{\boldsymbol{X}_1} d\lambda_{\boldsymbol{X}_2} \cdots d\lambda_{\boldsymbol{X}_{n-2}} -$$

$$\int_{\mathbb{R}^{p_{n-2}}} \cdots \int_{\mathbb{R}^{p_2}} \int_{\mathbb{R}^{p_1}} e^{i\boldsymbol{t}_1\cdot\boldsymbol{x}_1} e^{i\boldsymbol{t}_2\cdot\boldsymbol{x}_2} \cdots e^{i\boldsymbol{t}_{n-2}\cdot\boldsymbol{x}_{n-2}} \cdot d\lambda_{\boldsymbol{X}_1|\boldsymbol{x}_2\cdots\boldsymbol{x}_n} d\lambda_{\boldsymbol{X}_2|\boldsymbol{x}_3\cdots\boldsymbol{x}_n} \cdots d\lambda_{\boldsymbol{X}_{n-2}|\boldsymbol{x}_n\boldsymbol{x}_{n-1}}$$

Let $\boldsymbol{t}_i = \boldsymbol{0}$ for $i = 1, 2, \cdots, n-3$. Then you obtain

$$\int_{\mathbb{R}^{p_{n-2}}} e^{i\boldsymbol{t}_{n-2}\cdot\boldsymbol{x}_{n-2}} d\lambda_{\boldsymbol{X}_{n-2}} = \int_{\mathbb{R}^{p_{n-2}}} e^{i\boldsymbol{t}_{n-2}\cdot\boldsymbol{x}_{n-2}} d\lambda_{\boldsymbol{X}_{n-2}|\boldsymbol{x}_n\boldsymbol{x}_{n-1}}$$

and so $\lambda_{\boldsymbol{X}_{n-2}} = \lambda_{\boldsymbol{X}_{n-2}|\boldsymbol{x}_n\boldsymbol{x}_{n-1}}$ for $\boldsymbol{x}_{n-1}$ off a set of $\lambda_{\boldsymbol{X}_{n-1}}$ measure zero. Continuing this way, it follows that

$$\lambda_{\boldsymbol{X}_{n-k}} = \lambda_{\boldsymbol{X}_{n-k}|\boldsymbol{x}_n\boldsymbol{x}_{n-1}\cdots\boldsymbol{x}_{n-k+1}}$$

for $\boldsymbol{x}_{n-k+1}$ off a set of $\lambda_{\boldsymbol{X}_{n-k+1}}$ measure zero. Thus if E is Borel in $\mathbb{R}^{p_{n-1}} \times \cdots \times \mathbb{R}^{p_1}$,

$$\int_{\mathbb{R}^{p_n}\times\cdots\times\mathbb{R}^{p_1}} \mathcal{X}_E d\lambda_{(\boldsymbol{X}_1\cdots\boldsymbol{X}_n)} =$$

$$\int_{\mathbb{R}^{p_n}} \cdots \int_{\mathbb{R}^{p_2}} \int_{\mathbb{R}^{p_1}} \mathcal{X}_E d\lambda_{\boldsymbol{X}_1|\boldsymbol{x}_2\ldots\boldsymbol{x}_n} d\lambda_{\boldsymbol{X}_2|\boldsymbol{x}_3\cdots\boldsymbol{x}_n} \cdots d\lambda_{\boldsymbol{X}_{n-1}|\boldsymbol{x}_n} d\lambda_{\boldsymbol{X}_n}$$

$$\int_{\mathbb{R}^{p_n}} \cdots \int_{\mathbb{R}^{p_2}} \int_{\mathbb{R}^{p_1}} \mathcal{X}_E d\lambda_{\boldsymbol{X}_1|\boldsymbol{x}_2\ldots\boldsymbol{x}_n} d\lambda_{\boldsymbol{X}_2|\boldsymbol{x}_3\cdots\boldsymbol{x}_n} \cdots d\lambda_{\boldsymbol{X}_{n-1}} d\lambda_{\boldsymbol{X}_n}$$

$$\vdots$$

$$= \int_{\mathbb{R}^{p_n}} \cdots \int_{\mathbb{R}^{p_2}} \int_{\mathbb{R}^{p_1}} \mathcal{X}_E d\lambda_{\boldsymbol{X}_1} d\lambda_{\boldsymbol{X}_2} \cdots d\lambda_{\boldsymbol{X}_n}$$

One could achieve this iterated integral in any order by similar arguments to the above. By Definition 27.2.2 and the discussion which follows, this implies that the random variables $\boldsymbol{X}_i$ are independent. ■

Here is another proof of the Doob Dynkin lemma based on differentiation theory.

Lemma 27.4.3 *Suppose $\boldsymbol{X},\boldsymbol{Y}_1,\boldsymbol{Y}_2,\cdots,\boldsymbol{Y}_k$ are random vectors $\boldsymbol{X}$ having values in $\mathbb{R}^n$ and $\boldsymbol{Y}_j$ having values in $\mathbb{R}^{p_j}$ and*

$$\boldsymbol{X},\boldsymbol{Y}_j \in L^1(\Omega).$$

Suppose $\boldsymbol{X}$ is $\sigma(\boldsymbol{Y}_1,\cdots,\boldsymbol{Y}_k)$ measurable. Thus

$$\left\{\boldsymbol{X}^{-1}(E) : E \text{ Borel}\right\} \subseteq \left\{(\boldsymbol{Y}_1,\cdots,\boldsymbol{Y}_k)^{-1}(F) : F \text{ is Borel in } \prod_{j=1}^{k}\mathbb{R}^{p_j}\right\}$$

Then there exists a Borel function, $\boldsymbol{g} : \prod_{j=1}^{k}\mathbb{R}^{p_j} \to \mathbb{R}^n$ such that

$$\boldsymbol{X} = \boldsymbol{g}(\boldsymbol{Y}_1,\boldsymbol{Y}_2,\cdots,\boldsymbol{Y}_k).$$

Proof: For the sake of brevity, denote by $\boldsymbol{Y}$ the vector $(\boldsymbol{Y}_1,\cdots,\boldsymbol{Y}_k)$ and by $\boldsymbol{y}$ the vector $(\boldsymbol{y}_1,\cdots,\boldsymbol{y}_k)$ and let $\prod_{j=1}^{k}\mathbb{R}^{p_j} \equiv \mathbb{R}^P$. For E a Borel set of $\mathbb{R}^n$,

$$\begin{aligned}\int_{\boldsymbol{Y}^{-1}(E)} \boldsymbol{X}\,dP &= \int_{\mathbb{R}^n\times\mathbb{R}^P} \mathcal{X}_{\mathbb{R}^n\times E}(\boldsymbol{x},\boldsymbol{y})\,\boldsymbol{x}\,d\lambda_{(\boldsymbol{X},\boldsymbol{Y})} \\ &= \int_E \int_{\mathbb{R}^n} \boldsymbol{x}\,d\lambda_{\boldsymbol{X}|\boldsymbol{y}}\,d\lambda_{\boldsymbol{Y}}. \end{aligned} \tag{27.12}$$

Consider the function $\boldsymbol{y} \to \int_{\mathbb{R}^n} \boldsymbol{x}\,d\lambda_{\boldsymbol{X}|\boldsymbol{y}}$. Since $d\lambda_{\boldsymbol{Y}}$ is a Radon measure having inner and outer regularity, it follows the above function is equal to a Borel function for $\lambda_{\boldsymbol{Y}}$ a.e. $\boldsymbol{y}$. This function will be denoted by $\boldsymbol{g}$. Then from 27.12

$$\begin{aligned}\int_{\boldsymbol{Y}^{-1}(E)} \boldsymbol{X}\,dP &= \int_E \boldsymbol{g}(\boldsymbol{y})\,d\lambda_{\boldsymbol{Y}} = \int_{\mathbb{R}^P} \mathcal{X}_E(\boldsymbol{y})\,\boldsymbol{g}(\boldsymbol{y})\,d\lambda_{\boldsymbol{Y}} \\ &= \int_\Omega \mathcal{X}_E(\boldsymbol{Y}(\omega))\,\boldsymbol{g}(\boldsymbol{Y}(\omega))\,dP = \int_{\boldsymbol{Y}^{-1}(E)} \boldsymbol{g}(\boldsymbol{Y}(\omega))\,dP\end{aligned}$$

and since $\boldsymbol{Y}^{-1}(E)$ is an arbitrary element of $\sigma(\boldsymbol{Y})$, this shows that since $\boldsymbol{X}$ is $\sigma(\boldsymbol{Y})$ measurable, $\boldsymbol{X} = \boldsymbol{g}(\boldsymbol{Y})$ P a.e. ■

What about the case where $\boldsymbol{X}$ is not necessarily measurable in $\sigma(\boldsymbol{Y}_1,\cdots,\boldsymbol{Y}_k)$?

Lemma 27.4.4 *There exists a unique function $\boldsymbol{Z}(\omega)$ which satisfies*

$$\int_F \boldsymbol{X}(\omega)\,dP = \int_F \boldsymbol{Z}(\omega)\,dP$$

for all

$$F \in \sigma(\boldsymbol{Y}_1,\cdots,\boldsymbol{Y}_k)$$

such that $\boldsymbol{Z}$ is $\sigma(\boldsymbol{Y}_1,\cdots,\boldsymbol{Y}_k)$ measurable. It is denoted by

$$E(\boldsymbol{X}|\sigma(\boldsymbol{Y}_1,\cdots,\boldsymbol{Y}_k))$$

Proof: It is like the above. Letting E be a Borel set in $\mathbb{R}^p$,

$$\int_{\boldsymbol{Y}^{-1}(E)} \boldsymbol{X}\,dP = \int_{\mathbb{R}^n\times\mathbb{R}^P} \mathcal{X}_{\mathbb{R}^n\times E}(\boldsymbol{x},\boldsymbol{y})\,\boldsymbol{x}\,d\lambda_{(\boldsymbol{X},\boldsymbol{Y})} = \int_E \int_{\mathbb{R}^n} \boldsymbol{x}\,d\lambda_{\boldsymbol{X}|\boldsymbol{y}}\,d\lambda_{\boldsymbol{Y}}.$$

Now let $g(y) \equiv E(X|y_1,\cdots,y_k)$ be a Borel representative of $\int_{\mathbb{R}^n} x d\lambda_{X|y}$ It follows $\omega \to g(Y(\omega)) = E(X|Y_1(\omega),\cdots,Y_k(\omega))$ is $\sigma(Y_1,\cdots,Y_k)$ measurable because by definition $\omega \to Y(\omega)$ is $\sigma(Y_1,\cdots,Y_k)$ measurable and a Borel measurable function composed with a measurable one is still measurable. It follows that for all E Borel in $\mathbb{R}^p$,

$$\int_{Y^{-1}(E)} X dP = \int_E E(X|y_1,\cdots,y_k)\, d\lambda_Y$$

$$= \int_{Y^{-1}(E)} E(X|Y_1(\omega),\cdots,Y_k(\omega))\, dP$$

so $Z(\omega) = E(X|Y_1(\omega),\cdots,Y_k(\omega))$ works because a generic set of $\sigma(Y_1,\cdots,Y_k)$ is $Y^{-1}(E)$ for E a Borel set in $\mathbb{R}^p$. If both Z, Z_1 work, then for all

$$F \in \sigma(Y_1,\cdots,Y_k),$$

$$\int_F (Z - Z_1)\, dP = 0$$

Since F is arbitrary, some routine computations show $Z = Z_1$ a.e. ■

Observation 27.4.5 *Note that a.e.*

$$E(X|Y_1(\omega),\cdots,Y_k(\omega)) = E(X|\sigma(Y_1,\cdots,Y_k))$$

where the one on the left is the expected value of X given values of $Y_j(\omega)$. This one corresponds to the sort of thing we say in words. The one on the right is an abstract concept which is usually obtained using the Radon Nikodym theorem and its description is given in the lemma. This lemma shows that its meaning is really to take the expected value of X given values for the Y_k.

27.5 Characteristic Functions for Measures

Recall the characteristic function for a random variable having values in $\mathbb{R}^p$. I will give a review of this to begin with. Then the concept will be generalized to random variables (vectors) which have values in a real separable Banach space.

Definition 27.5.1 *Let X be a random variable. The characteristic function is*

$$\phi_X(t) \equiv E\left(e^{it\cdot X}\right) \equiv \int_\Omega e^{it\cdot X(\omega)} dP = \int_{\mathbb{R}^p} e^{it\cdot x} d\lambda_X$$

the last equation holding by Proposition 26.1.12 on Page 745.

Recall the following fundamental lemma and definition, Lemma 13.2.4 on Page 387.

Definition 27.5.2 *For $T \in \mathcal{G}^*$, define $FT, F^{-1}T \in \mathcal{G}^*$ by*

$$FT(\phi) \equiv T(F\phi),\ F^{-1}T(\phi) \equiv T\left(F^{-1}\phi\right)$$

Lemma 27.5.3 *F and F^{-1} are both one to one, onto, and are inverses of each other.*

The main result on characteristic functions is the following in Theorem 27.1.4 on Page 763 which is stated here for convenience.

Theorem 27.5.4 *Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be random vectors with values in $\mathbb{R}^p$ and suppose $E\left(e^{i\boldsymbol{t}\cdot\boldsymbol{X}}\right) = E\left(e^{i\boldsymbol{t}\cdot\boldsymbol{Y}}\right)$ for all $\boldsymbol{t}\in\mathbb{R}^p$. Then $\lambda_{\boldsymbol{X}} = \lambda_{\boldsymbol{Y}}$.*

I want to do something similar for random variables which have values in a separable real Banach space, E instead of $\mathbb{R}^p$.

Corollary 27.5.5 *Let $\mathscr{K}$ be a π system of subsets of Ω and suppose two probability measures, μ and ν defined on $\sigma(\mathscr{K})$ are equal on $\mathscr{K}$. Then $\mu = \nu$.*

Proof: This follows from the Lemma 9.3.2 on Page 247. Let

$$\mathscr{G} \equiv \{E \in \sigma(\mathscr{K}) : \mu(E) = \nu(E)\}$$

Then $\mathscr{K} \subseteq \mathscr{G}$, since μ and ν are both probability measures, it follows that if $E \in \mathscr{G}$, then so is E^C. Since these are measures, if $\{A_i\}$ is a sequence of disjoint sets from $\mathscr{G}$ then

$$\mu(\cup_{i=1}^{\infty} A_i) = \sum_i \mu(A_i) = \sum_i \nu(A_i) = \nu(\cup_{i=1}^{\infty} A)$$

and so from Lemma 9.3.2, $\mathscr{G} = \sigma(\mathscr{K})$. ■

Next recall the following fundamental lemma used to prove Pettis' theorem. It is proved on Page 671 but is stated here for convenience.

Lemma 27.5.6 *If E is a separable Banach space with B' the closed unit ball in E', then there exists a sequence $\{f_n\}_{n=1}^{\infty} \equiv D' \subseteq B'$ with the property that for every $x \in E$,*

$$\|x\| = \sup_{f\in D'} |f(x)|$$

Definition 27.5.7 *Let E be a separable real Banach space. A cylindrical set is one which is of the form*

$$\{x \in E : x_i^*(x) \in \Gamma_i, i = 1,2,\cdots,m\}$$

where here $x_i^* \in E'$ and Γ_i is a Borel set in $\mathbb{R}$.

It is obvious that $\emptyset$ is a cylindrical set and that the intersection of two cylindrical sets is another cylindrical set. Thus the cylindrical sets form a π system. What is the smallest σ algebra containing the cylindrical sets? It is the Borel sets of E. This is a special case of Lemma 26.5.2. Recall why this was. Letting $\{f_n\}_{n=1}^{\infty} = D'$ be the sequence of Lemma 27.5.6 it follows that

$$\begin{aligned} & \{x \in E : \|x-a\| \le \delta\} \\ = & \left\{x \in E : \sup_{f\in D'} |f(x-a)| \le \delta\right\} = \left\{x \in E : \sup_{f\in D'} |f(x) - f(a)| \le \delta\right\} \\ = & \cap_{n=1}^{\infty} \left\{x \in E : f_n(x) \in \overline{B(f_n(a),\delta)}\right\} \end{aligned}$$

which yields a countable intersection of cylindrical sets. It follows the smallest σ algebra containing the cylindrical sets contains the closed balls and hence the open balls and consequently the open sets and so it contains the Borel sets. However, each cylindrical set is a Borel set and so in fact this σ algebra equals $\mathscr{B}(E)$.

From Corollary 27.5.5 it follows that two probability measures which are equal on the cylindrical sets are equal on the Borel sets $\mathscr{B}(E)$.

Definition 27.5.8 *Let μ be a probability measure on $\mathscr{B}(E)$ where E is a real separable Banach space. Then for $x^* \in E'$,*

$$\phi_\mu(x^*) \equiv \int_E e^{ix^*(x)} d\mu(x).$$

ϕ_μ is called the characteristic function for the measure μ.

Note this is a little different than earlier when the symbol $\phi_X(t)$ was used and X was a random variable. Here the focus is more on the measure than a random variable X such that its distribution measure is μ. It might appear this is a more general concept but in fact this is not the case. You could just consider the separable Banach space or Polish space with the Borel σ algebra as your probabililty space and then consider the identity map as a random variable having the given measure as a distribution measure. Of course a major result is the one which says that the characteristic function determines the measures.

Theorem 27.5.9 *Let μ and ν be two probability measures on $\mathscr{B}(E)$ where E is a separable real Banach space. Suppose*

$$\phi_\mu(x^*) = \phi_\nu(x^*)$$

for all $x^* \in E'$. Then $\mu = \nu$.

Proof: It suffices to verify that $\mu(A) = \nu(A)$ for all $A \in \mathscr{K}$ where $\mathscr{K}$ is the set of cylindrical sets. Fix $g_n \in (E')^n$. Thus the two measures are equal if for all such g_n, $n \in \mathbb{N}$,

$$\mu\left(g_n^{-1}(B)\right) = \nu\left(g_n^{-1}(B)\right)$$

for B a Borel set in $\mathbb{R}^n$. Of course, for such a choice of $g_n \in (E')^n$, there are measures defined on the Borel sets of $\mathbb{R}^n$ μ_n and ν_n which are given by

$$\mu_n(B) \equiv \mu\left(g_n^{-1}(B)\right),\ \nu_n(B) \equiv \nu\left(g_n^{-1}(B)\right)$$

and so it suffices to verify that these two measures are equal. So what are their characteristic functions? Note that g_n is a random variable taking E to $\mathbb{R}^n$ and μ_n, ν_n are just the probability distribution measures of this random variable. Therefore,

$$\phi_{\mu_n}(t) \equiv \int_{\mathbb{R}^n} e^{it\cdot s} d\mu_n = \int_E e^{it\cdot g_n(x)} d\mu$$

Similarly,

$$\phi_{\nu_n}(t) \equiv \int_{\mathbb{R}^n} e^{it\cdot s} d\nu_n = \int_E e^{it\cdot g_n(x)} d\nu$$

Now $t\cdot g_n \in E'$ and so by assumption, the two ends of the above are equal. Hence $\phi_{\mu_n}(t) = \phi_{\nu_n}(t)$ and so by Theorem 27.1.6, $\mu_n = \nu_n$ which, as shown above, implies $\mu = \nu$. ■

27.6 Independence in Banach Space

I will consider the relation between the characteristic function and independence of random variables having values in a Banach space. Recall an earlier proposition which relates independence of random vectors with characteristic functions. It is proved starting on Page 773.

Proposition 27.6.1 *Let $\{\boldsymbol{X}_k\}_{k=1}^n$ be random vectors such that $\boldsymbol{X}_k$ has values in $\mathbb{R}^{p_k}$. Then the random vectors are independent if and only if*

$$E\left(e^{iP}\right) = \prod_{j=1}^{n} E\left(e^{i\boldsymbol{t}_j\cdot\boldsymbol{X}_j}\right)$$

where $P \equiv \sum_{j=1}^n \boldsymbol{t}_j \cdot \boldsymbol{X}_j$ for $\boldsymbol{t}_j \in \mathbb{R}^{p_j}$.

It turns out there is a generalization of the above proposition to the case where the random variables have values in a real separable Banach space. Before proving this recall an earlier theorem which had to do with reducing to the case where the random variables had values in $\mathbb{R}^n$, Theorem 26.6.1. It is restated here for convenience.

Theorem 27.6.2 *The random variables $\{X_i\}_{i\in I}$ are independent if whenever*

$$\{i_1,\cdots,i_n\} \subseteq I,$$

$m_{i_1},\cdots,m_{i_n}$ are positive integers, and $\boldsymbol{g}_{m_{i_1}},\cdots,\boldsymbol{g}_{m_{i_n}}$ are in

$$\left(E'\right)^{m_{i_1}},\cdots,\left(E'\right)^{m_{i_n}}$$

respectively, $\left\{\boldsymbol{g}_{m_{i_j}}\circ X_{i_j}\right\}_{j=1}^n$ are independent random vectors having values in

$$\mathbb{R}^{m_{i_1}},\cdots,\mathbb{R}^{m_{i_n}}$$

respectively.

Now here is the theorem about independence and the characteristic functions.

Theorem 27.6.3 *Let $\{X_k\}_{k=1}^n$ be random variables such that X_k has values in E_k, a real separable Banach space. Then the random variables are independent if and only if*

$$E\left(e^{iP}\right) = \prod_{j=1}^{n} E\left(e^{it_j^*(X_j)}\right)$$

where $P \equiv \sum_{j=1}^n t_j^*(X_j)$ for $t_j^* \in E_j'$.

Proof: If the random variables are independent, then so are the random variables, $t_j^*(X_j)$ and so the equation follows.

The interesting case is when the equation holds.

It suffices to consider only the case where each $E_k = E$. This is because you can consider each X_j to have values in $\prod_{k=1}^n E_k$ by letting X_j take its values in the j^{th} component of the product and 0 in the other components. Can you draw the conclusion the random variables are independent? By Theorem 26.6.1, it suffices to show the random variables $\left\{\boldsymbol{g}_{m_k}\circ X_k\right\}_{k=1}^n$ are independent where $\boldsymbol{g}_{m_k} = \left(x_1^*,\cdots,x_{m_k}^*\right) \in (E')^{m_k}$. This happens if whenever $\boldsymbol{t}_{m_k} \in \mathbb{R}^{m_k}$ and

$$P = \sum_{k=1}^{n} \boldsymbol{t}_{m_k}\cdot\left(\boldsymbol{g}_{m_k}\circ X_k\right),$$

it follows

$$E\left(e^{iP}\right)=\prod_{k=1}^{n}E\left(e^{i\boldsymbol{t}_{m_k}\cdot\left(\boldsymbol{g}_{m_k}\circ X_k\right)}\right). \tag{27.13}$$

However, the expression on the right in 27.13 equals

$$\prod_{k=1}^{n}E\left(e^{i\left(\boldsymbol{t}_{m_k}\cdot\boldsymbol{g}_{m_k}\right)\circ X_k}\right)$$

and $\boldsymbol{t}_{m_k}\cdot\boldsymbol{g}_{m_k}\equiv\sum_{j=1}^{m_k}t_jx_j^*\in E'$. Also the expression on the left equals

$$E\left(e^{i\sum_{k=1}^{n}\boldsymbol{t}_{m_k}\cdot\boldsymbol{g}_{m_k}\circ X_k}\right)$$

Therefore, by assumption, 27.13 holds. ■

There is an obvious corollary which is useful.

Corollary 27.6.4 *Let $\{X_k\}_{k=1}^n$ be random variables such that X_k has values in E_k, a real separable Banach space. Then the random variables are independent if and only if*

$$E\left(e^{iP}\right)=\prod_{j=1}^{n}E\left(e^{it_j^*(X_j)}\right)$$

where $P\equiv\sum_{j=1}^{n}t_j^*(X_j)$ for $t_j^*\in M_j$ where M_j is a dense subset of E_j'.

Proof: The easy direction follows from Theorem 27.6.3. Suppose then the above equation holds for all $t_j^*\in M_j$. Then let $t_j^*\in E'$ and let $\left\{t_{nj}^*\right\}$ be a sequence in M_j such that $\lim_{n\to\infty}t_{nj}^*=t_j^*$ in E'. Then define

$$P\equiv\sum_{j=1}^{n}t_j^*X_j,\ P_n\equiv\sum_{j=1}^{n}t_{nj}^*X_j.$$

It follows

$$E\left(e^{iP}\right)=\lim_{n\to\infty}E\left(e^{iP_n}\right)=\lim_{n\to\infty}\prod_{j=1}^{n}E\left(e^{it_{nj}^*(X_j)}\right)=\prod_{j=1}^{n}E\left(e^{it_j^*(X_j)}\right)\ ■$$

27.7 Convolution and Sums

Lemma 26.1.9 on Page 745 makes possible a definition of convolution of two probability measures defined on $\mathscr{B}(E)$ where E is a separable Banach space. I will first show a little theorem about density of continuous functions in $L^p(E)$ and then define the convolution of two finite measures. First here is a simple technical lemma.

Lemma 27.7.1 *Suppose K is a compact subset of U an open set in E a metric space. Then there exists $\delta>0$ such that*

$$\operatorname{dist}(x,K)+\operatorname{dist}\left(x,U^C\right)\geq\delta \text{ for all } x\in E.$$

Proof: For each $x \in K$, there exists a ball, $B(x,\delta_x)$ such that $B(x,3\delta_x) \subseteq U$. Finitely many of these balls cover K because K is compact, say $\{B(x_i,\delta_{x_i})\}_{i=1}^m$. Let

$$0 < \delta < \min(\delta_{x_i} : i = 1,2,\cdots,m).$$

Now pick any $x \in K$. Then $x \in B(x_i,\delta_{x_i})$ for some x_i and so

$$B(x,\delta) \subseteq B(x_i,2\delta_{x_i}) \subseteq U.$$

Therefore, for any $x \in K$, $\operatorname{dist}(x,U^C) \geq \delta$. If $x \in B(x_i,2\delta_{x_i})$ for some x_i, it follows that $\operatorname{dist}(x,U^C) \geq \delta$ because then $B(x,\delta) \subseteq B(x_i,3\delta_{x_i}) \subseteq U$. If $x \notin B(x_i,2\delta_{x_i})$ for any of the x_i, then $x \notin B(y,\delta)$ for any $y \in K$ because all these sets are contained in some $B(x_i,2\delta_{x_i})$. Consequently $\operatorname{dist}(x,K) \geq \delta$. ■

From this lemma, there is an easy corollary.

Corollary 27.7.2 *Suppose K is a compact subset of U, an open set in E a metric space. Then there exists a uniformly continuous function f defined on all of E, having values in $[0,1]$ such that $f(x) = 0$ if $x \notin U$ and $f(x) = 1$ if $x \in K$.*

Proof: Consider

$$f(x) \equiv \frac{\operatorname{dist}(x,U^C)}{\operatorname{dist}(x,U^C) + \operatorname{dist}(x,K)}.$$

Then some algebra yields

$$\left|f(x) - f(x')\right| \leq \frac{1}{\delta}\left(\left|\operatorname{dist}(x,U^C) - \operatorname{dist}(x',U^C)\right| + \left|\operatorname{dist}(x,K) - \operatorname{dist}(x',K)\right|\right)$$

where δ is the constant of Lemma 27.7.1. Now it is a general fact that

$$\left|\operatorname{dist}(x,S) - \operatorname{dist}(x',S)\right| \leq d(x,x').$$

See Proposition 3.6.6. Therefore, $|f(x) - f(x')| \leq \frac{2}{\delta} d(x,x')$ and this proves the corollary. ■

Now suppose μ is a finite measure defined on the Borel sets of a separable Banach space E. It was shown above that μ is inner and outer regular. Lemma 26.1.9 on Page 745 shows that μ is inner regular in the usual sense with respect to compact sets. This makes possible the following theorem.

Theorem 27.7.3 *Let μ be a finite measure on $\mathcal{B}(E)$ where E is a separable Banach space and let $f \in L^p(E;\mu)$. Then for any $\varepsilon > 0$, there exists a uniformly continuous, bounded g defined on E such that*

$$\|f - g\|_{L^p(E)} < \varepsilon.$$

Proof: As usual in such situations, it suffices to consider only $f \geq 0$. Then by Theorem 9.1.6 on Page 244 and an application of the monotone convergence theorem, there exists a simple measurable function,

$$s(x) \equiv \sum_{k=1}^m c_k \mathcal{X}_{A_k}(x)$$

such that $||f - s||_{L^p(E)} < \varepsilon/2$. Now by regularity of μ there exist compact sets, K_k and open sets, V_k such that $2\sum_{k=1}^m |c_k| \mu(V_k \setminus K)^{1/p} < \varepsilon/2$ and by Corollary 27.7.2 there exist

uniformly continuous functions g_k having values in $[0,1]$ such that $g_k = 1$ on K_k and 0 on V_k^C. Then consider

$$g(x) = \sum_{k=1}^{m} c_k g_k(x).$$

This function is bounded and uniformly continuous. Furthermore,

$$\|s-g\|_{L^p(E)} \leq \left(\int_E \left| \sum_{k=1}^{m} c_k \mathcal{X}_{A_k}(x) - \sum_{k=1}^{m} c_k g_k(x) \right|^p d\mu \right)^{1/p}$$

$$\leq \left(\int_E \left(\sum_{k=1}^{m} |c_k| \left| \mathcal{X}_{A_k}(x) - g_k(x) \right| \right)^p \right)^{1/p} \leq \sum_{k=1}^{m} |c_k| \left(\int_E \left| \mathcal{X}_{A_k}(x) - g_k(x) \right|^p d\mu \right)^{1/p}$$

$$\leq \sum_{k=1}^{m} |c_k| \left(\int_{V_k \setminus K_k} 2^p d\mu \right)^{1/p} = 2 \sum_{k=1}^{m} |c_k| \mu(V_k \setminus K)^{1/p} < \varepsilon/2.$$

Therefore,

$$\|f-g\|_{L^p} \leq \|f-s\|_{L^p} + \|s-g\|_{L^p} < \varepsilon/2 + \varepsilon/2 \ \blacksquare$$

Lemma 27.7.4 *Let $A \in \mathcal{B}(E)$ where μ is a finite measure on $\mathcal{B}(E)$ for E a separable Banach space. Also let $x_i \in E$ for $i = 1, 2, \cdots, m$. Then for $\boldsymbol{x} \in E^m$,*

$$\boldsymbol{x} \to \mu \left(A + \sum_{i=1}^{m} x_i \right), \ \boldsymbol{x} \to \mu \left(A - \sum_{i=1}^{m} x_i \right)$$

are Borel measurable functions. Furthermore, the above functions are

$$\mathcal{B}(E) \times \cdots \times \mathcal{B}(E)$$

measurable where the above denotes the product measurable sets as described in Theorem 10.14.8 on Page 314.

Proof: First consider the case where $A = U$, an open set. Let

$$\boldsymbol{y} \in \left\{ \boldsymbol{x} \in E^m : \mu \left(U + \sum_{i=1}^{m} x_i \right) > \alpha \right\} \tag{27.14}$$

Then from Lemma 26.1.9 on Page 745 there exists a compact set, $K \subseteq U + \sum_{i=1}^{m} y_i$ such that $\mu(K) > \alpha$. Then if $\boldsymbol{y}'$ is close enough to $\boldsymbol{y}$, it follows $K \subseteq U + \sum_{i=1}^{m} y_i'$ also. Therefore, for all $\boldsymbol{y}'$ close enough to $\boldsymbol{y}$,

$$\mu \left(U + \sum_{i=1}^{m} y_i' \right) \geq \mu(K) > \alpha.$$

In other words the set described in 27.14 is an open set and so $\boldsymbol{y} \to \mu(U + \sum_{i=1}^{m} y_i)$ is Borel measurable whenever U is an open set in E.

Define a π system, $\mathcal{K}$ to consist of all open sets in E. Then define $\mathcal{G}$ as

$$\left\{ A \in \sigma(\mathcal{K}) = \mathcal{B}(E) : \boldsymbol{y} \to \mu \left(A + \sum_{i=1}^{m} y_i \right) \text{ is Borel measurable} \right\}$$

I just showed $\mathscr{G} \supseteq \mathscr{K}$. Now suppose $A \in \mathscr{G}$. Then

$$\mu\left(A^C + \sum_{i=1}^m y_i\right) = \mu(E) - \mu\left(A + \sum_{i=1}^m y_i\right)$$

and so $A^C \in \mathscr{G}$ whenever $A \in \mathscr{G}$. Next suppose $\{A_i\}$ is a sequence of disjoint sets of $\mathscr{G}$. Then

$$\mu\left((\cup_{i=1}^\infty A_i) + \sum_{j=1}^m y_j\right) = \mu\left(\cup_{i=1}^\infty \left(A_i + \sum_{j=1}^m y_j\right)\right) = \sum_{i=1}^\infty \mu\left(A_i + \sum_{j=1}^m y_j\right)$$

and so $\cup_{i=1}^\infty A_i \in \mathscr{G}$ because the above is the sum of Borel measurable functions. By the lemma on π systems, Lemma 9.3.2 on Page 247, it follows $\mathscr{G} = \sigma(\mathscr{K}) = \mathscr{B}(E)$. Similarly, $x \to \mu\left(A - \sum_{j=1}^m x_j\right)$ is also Borel measurable whenever $A \in \mathscr{B}(E)$. Finally note that

$$\mathscr{B}(E) \times \cdots \times \mathscr{B}(E)$$

contains the open sets of E^m because the separability of E implies the existence of a countable basis for the topology of E^m consisting of sets of the form $\prod_{i=1}^m U_i$ where the U_i come from a countable basis for E. Since every open set is the countable union of sets like the above, each being a measurable box, the open sets are contained in $\mathscr{B}(E) \times \cdots \times \mathscr{B}(E)$ which implies $\mathscr{B}(E^m) \subseteq \mathscr{B}(E) \times \cdots \times \mathscr{B}(E)$ also. ■

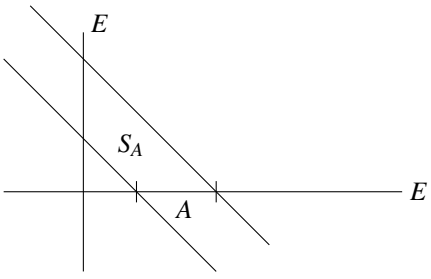
With this lemma, it is possible to define the convolution of two finite measures.

Definition 27.7.5 *Let μ and ν be two finite measures on $\mathscr{B}(E)$, for E a separable Banach space. Then define a new measure, $\mu * \nu$ on $\mathscr{B}(E)$ as follows*

$$\mu * \nu(A) \equiv \int_E \nu(A - x)\, d\mu(x).$$

This is well defined because of Lemma 27.7.4 which says that $x \to \nu(A - x)$ is Borel measurable.

Here is an interesting theorem about convolutions. However, first here is a little lemma. The following picture is descriptive of the set described in the following lemma.



Lemma 27.7.6 *For A a Borel set in E, a separable Banach space, define*

$$S_A \equiv \{(x,y) \in E \times E : x + y \in A\}$$

Then $S_A \in \mathscr{B}(E) \times \mathscr{B}(E)$, the σ algebra of product measurable sets, the smallest σ algebra which contains all the sets of the form $A \times B$ where A and B are Borel.

Proof: Let $\mathscr{K}$ denote the open sets in E. Then $\mathscr{K}$ is a π system. Let

$$\mathscr{G} \equiv \{A \in \sigma(\mathscr{K}) = \mathscr{B}(E) : S_A \in \mathscr{B}(E) \times \mathscr{B}(E)\}.$$

Then $\mathscr{K} \subseteq \mathscr{G}$ because if $U \in \mathscr{K}$ then S_U is an open set in $E \times E$ and all open sets are in $\mathscr{B}(E) \times \mathscr{B}(E)$ because a countable basis for the topology of $E \times E$ are sets of the form $B \times C$ where B and C come from a countable basis for E. Therefore, $\mathscr{K} \subseteq \mathscr{G}$. Now let $A \in \mathscr{G}$. For $(x,y) \in E \times E$, either $x+y \in A$ or $x+y \notin A$. Hence $E \times E = S_A \cup S_{A^C}$ which shows that if $A \in \mathscr{G}$ then so is A^C. Finally if $\{A_i\}$ is a sequence of disjoint sets of $\mathscr{G}$

$$S_{\cup_{i=1}^{\infty} A_i} = \cup_{i=1}^{\infty} S_{A_i}$$

and this shows that $\mathscr{G}$ is also closed with respect to countable unions of disjoint sets. Therefore, by the lemma on π systems, Lemma 9.3.2 on Page 247 it follows $\mathscr{G} = \sigma(\mathscr{K}) = \mathscr{B}(E)$. This proves the lemma.

Theorem 27.7.7 *Let μ, ν, and λ be finite measures on $\mathscr{B}(E)$ for E a separable Banach space. Then*

$$\mu * \nu = \nu * \mu \tag{27.15}$$

$$(\mu * \nu) * \lambda = \mu * (\nu * \lambda) \tag{27.16}$$

If μ is the distribution for an E valued random variable, X and if ν is the distribution for an E valued random variable, Y, and X and Y are independent, then $\mu * \nu$ is the distribution for the random variable, $X+Y$. Also the characteristic function of a convolution equals the product of the characteristic functions.

Proof: First consider 27.15. Letting $A \in \mathscr{B}(E)$, the following computation holds from Fubini's theorem and Lemma 27.7.6

$$\begin{aligned} \mu * \nu(A) &\equiv \int_E \nu(A-x)\, d\mu(x) = \int_E \int_E \mathscr{X}_{S_A}(x,y)\, d\nu(y)\, d\mu(x) \\ &= \int_E \int_E \mathscr{X}_{S_A}(x,y)\, d\mu(x)\, d\nu(y) = \nu * \mu(A). \end{aligned}$$

Next consider 27.16. Using 27.15 whenever convenient,

$$\begin{aligned} (\mu * \nu) * \lambda(A) &\equiv \int_E (\mu * \nu)(A-x)\, d\lambda(x) \\ &= \int_E \int_E \nu(A-x-y)\, d\mu(y)\, d\lambda(x) \end{aligned}$$

while

$$\begin{aligned} \mu * (\nu * \lambda)(A) &\equiv \int_E (\nu * \lambda)(A-y)\, d\mu(y) \\ &= \int_E \int_E \nu(A-y-x)\, d\lambda(x)\, d\mu(y) \\ &= \int_E \int_E \nu(A-y-x)\, d\mu(y)\, d\lambda(x). \end{aligned}$$

The necessary product measurability comes from Lemma 27.7.4.

Recall

$$(\mu * \nu)(A) \equiv \int_E \nu(A-x)\, d\mu(x).$$

Therefore, if s is a simple function, $s(x) = \sum_{k=1}^n c_k \mathcal{X}_{A_k}(x)$,

$$\begin{aligned}\int_E s d(\mu * \nu) &= \sum_{k=1}^n c_k \int_E \nu(A_k - x)\, d\mu(x) = \int_E \sum_{k=1}^n c_k \nu(A_k - x)\, d\mu(x) \\ &= \int_E \sum_{k=1}^n c_k \mathcal{X}_{A_k - x}(y)\, d\nu(y)\, d\mu(x) = \int_E \int_E s(x+y)\, d\nu(y)\, d\mu(x)\end{aligned}$$

Approximating with simple functions it follows that whenever f is bounded and measurable or nonnegative and measurable,

$$\int_E f d(\mu * \nu) = \int_E \int_E f(x+y)\, d\nu(y)\, d\mu(x) \tag{27.17}$$

Therefore, letting $Z = X + Y$, and λ the distribution of Z, it follows from independence of X and Y that for $t^* \in E'$,

$$\phi_\lambda(t^*) \equiv E\left(e^{it^*(Z)}\right) = E\left(e^{it^*(X+Y)}\right) = E\left(e^{it^*(X)}\right) E\left(e^{it^*(Y)}\right)$$

But also, it follows from 27.17

$$\begin{aligned}\phi_{(\mu*\nu)}(t^*) &= \int_E e^{it^*(z)} d(\mu * \nu)(z) = \int_E \int_E e^{it^*(x+y)} d\nu(y)\, d\mu(x) \\ &= \int_E \int_E e^{it^*(x)} e^{it^*(y)} d\nu(y)\, d\mu(x) \\ &= \left(\int_E e^{it^*(y)} d\nu(y)\right)\left(\int_E e^{it^*(x)} d\mu(x)\right) \\ &= E\left(e^{it^*(X)}\right) E\left(e^{it^*(Y)}\right)\end{aligned}$$

Since $\phi_\lambda(t^*) = \phi_{(\mu*\nu)}(t^*)$, it follows $\lambda = \mu * \nu$.

Note the last part of this argument shows the characteristic function of a convolution equals the product of the characteristic functions. ■

Chapter 28

The Normal Distribution

This particular distribution is likely the most important one in statistics and it will be essential to understand in developing the Wiener process later. To begin with, $\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{-\frac{1}{2}u^2}du = 1$ as is easily shown as done earlier by the standard calculus trick of

$$I = \int_{\mathbb{R}} e^{-\frac{1}{2}u^2}du, I^2 = \int_{\mathbb{R}}\int_{\mathbb{R}} e^{-\frac{1}{2}\left(u^2+v^2\right)}dudv$$

and then changing to polar coordinates to obtain $I^2 = 2\pi$. I will use this identity whenever convenient. Also useful is the following lemma.

Lemma 28.0.1 $\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{-\frac{1}{2}(u-it)^2}du = 1$.

Proof: $e^{-\frac{1}{2}(u-it)^2} = e^{-\frac{1}{2}\left(u^2-2itu-t^2\right)} = e^{-\frac{1}{2}u^2}e^{\frac{1}{2}t^2}\left(\cos(tu)+i\sin(tu)\right)$ and so, the integral equals

$$\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{-\frac{1}{2}u^2}e^{\frac{1}{2}t^2}\cos(tu)\,du$$

Now let $f(t) \equiv \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{-\frac{1}{2}u^2}e^{\frac{1}{2}t^2}\cos(tu)\,du$. Using the dominated convergence theorem,

$$f'(t) = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \frac{d}{dt}\left(e^{-\frac{1}{2}u^2}e^{\frac{1}{2}t^2}\cos(tu)\right)du$$

$$= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \left(e^{-\frac{1}{2}u^2}\left(te^{\frac{1}{2}t^2}\cos(tu) - e^{\frac{1}{2}t^2}u\sin(tu)\right)\right)du$$

Now $f(0)$ is known to be 1. Assume then that $t \neq 0$.

$$-\frac{1}{\sqrt{2\pi}}e^{\frac{1}{2}t^2}\int_{\mathbb{R}} e^{-\frac{1}{2}u^2}u\sin(tu)\,du = \frac{1}{\sqrt{2\pi}}e^{\frac{1}{2}t^2}\int_{\mathbb{R}} e^{-\frac{1}{2}u^2}t\cos(tu)\,du$$

and this shows that $f'(t) = 0$ so $f(t)$ is the constant 1. ■

28.1 The Multivariate Normal Distribution

The multivariate normal distribution is very important in statistics and it will be shown in this chapter why this is the case.

Definition 28.1.1 *A random vector $\boldsymbol{X}$, with values in $\mathbb{R}^p$ has a multivariate normal distribution written as $\boldsymbol{X} \sim N_p(\boldsymbol{m},\Sigma)$ if for all Borel $E \subseteq \mathbb{R}^p$,*

$$\lambda_{\boldsymbol{X}}(E) = \int_{\mathbb{R}^p} \mathcal{X}_E(\boldsymbol{x})\frac{1}{(2\pi)^{p/2}\det(\Sigma)^{1/2}}e^{\frac{-1}{2}(\boldsymbol{x}-\boldsymbol{m})^*\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{m})}dx$$

for μ a given vector and Σ a given positive definite symmetric matrix, called the covariance matrix. In case $p = 1$, this is called the variance.

Theorem 28.1.2 *For $\boldsymbol{X} \sim N_p(\boldsymbol{m},\Sigma), \boldsymbol{m} = E(\boldsymbol{X})$ and*

$$\Sigma = E\left((\boldsymbol{X}-\boldsymbol{m})(\boldsymbol{X}-\boldsymbol{m})^*\right).$$

Proof: Let R be an orthogonal transformation such that

$$R\Sigma R^* = D = \text{diag}\left(\sigma_1^2, \cdots, \sigma_p^2\right),\ \sigma_i > 0.$$

Changing the variable by $\boldsymbol{x} - \boldsymbol{m} = R^*\boldsymbol{y}$,

$$\begin{aligned} E(\boldsymbol{X}) &\equiv \int_{\mathbb{R}^p} \boldsymbol{x} e^{\frac{-1}{2}(\boldsymbol{x}-\boldsymbol{m})^*\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{m})} dx \left(\frac{1}{(2\pi)^{p/2}\det(\Sigma)^{1/2}}\right) \\ &= \int_{\mathbb{R}^p} (R^*\boldsymbol{y} + \boldsymbol{m}) e^{-\frac{1}{2}\boldsymbol{y}^* D^{-1}\boldsymbol{y}} dy \left(\frac{1}{(2\pi)^{p/2}\prod_{i=1}^p \sigma_i}\right) \\ &= \boldsymbol{m}\int_{\mathbb{R}^p} e^{-\frac{1}{2}\boldsymbol{y}^* D^{-1}\boldsymbol{y}} dy \left(\frac{1}{(2\pi)^{p/2}\prod_{i=1}^p \sigma_i}\right) = \boldsymbol{m} \end{aligned}$$

by Fubini's theorem and the easy to establish formula $\frac{1}{\sqrt{2\pi}\sigma}\int_{\mathbb{R}} e^{-\frac{y^2}{2\sigma^2}} dy = 1$, (let $u = y/\sigma$),

Next let

$$M \equiv E\left((\boldsymbol{X} - \boldsymbol{m})(\boldsymbol{X} - \boldsymbol{m})^*\right)$$

Thus, changing the variable as above by $\boldsymbol{x} - \boldsymbol{m} = R^*\boldsymbol{y}$

$$\begin{aligned} M &= \int_{\mathbb{R}^p} (\boldsymbol{x} - \boldsymbol{m})(\boldsymbol{x} - \boldsymbol{m})^* e^{\frac{-1}{2}(\boldsymbol{x}-\boldsymbol{m})^*\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{m})} dx \left(\frac{1}{(2\pi)^{p/2}\det(\Sigma)^{1/2}}\right) \\ &= R^*\int_{\mathbb{R}^p} \boldsymbol{y}\boldsymbol{y}^* e^{-\frac{1}{2}\boldsymbol{y}^* D^{-1}\boldsymbol{y}} dy \left(\frac{1}{(2\pi)^{p/2}\prod_{j=1}^p \sigma_j}\right) R \end{aligned}$$

If $i \neq j$, $(RMR^*)_{ij} = \int_{\mathbb{R}^p} y_i y_j e^{-\frac{1}{2}\boldsymbol{y}^* D^{-1}\boldsymbol{y}} dy \left(\frac{1}{(2\pi)^{p/2}\prod_{k=1}^p \sigma_k}\right) = 0$ so RMR^* is a diagonal matrix.

$$(RMR^*)_{ii} = \int_{\mathbb{R}^p} y_i^2 e^{-\frac{1}{2}\boldsymbol{y}^* D^{-1}\boldsymbol{y}} dy \left(\frac{1}{(2\pi)^{p/2}\prod_{j=1}^p \sigma_j}\right).$$

Using Fubini's theorem and the easy to establish equations,

$$\frac{1}{\sqrt{2\pi}\sigma}\int_{\mathbb{R}} e^{-\frac{y^2}{2\sigma^2}} dy = 1,\ \frac{1}{\sqrt{2\pi}\sigma}\int_{\mathbb{R}} y^2 e^{-\frac{y^2}{2\sigma^2}} dy = \sigma^2,$$

it follows $(RMR^*)_{ii} = \sigma_i^2$. Hence $RMR^* = D$ and so $M = R^*DR = \Sigma$. ■

Theorem 28.1.3 *Suppose $\boldsymbol{X}_1 \sim N_p(\boldsymbol{m}_1, \Sigma_1)$, $\boldsymbol{X}_2 \sim N_p(\boldsymbol{m}_2, \Sigma_2)$ and the two random vectors are independent. Then*

$$\boldsymbol{X}_1 + \boldsymbol{X}_2 \sim N_p(\boldsymbol{m}_1 + \boldsymbol{m}_2, \Sigma_1 + \Sigma_2). \tag{28.1}$$

Also, if $\boldsymbol{X} \sim N_p(\boldsymbol{m}, \Sigma)$ then $-\boldsymbol{X} \sim N_p(-\boldsymbol{m}, \Sigma)$. Furthermore, if $\boldsymbol{X} \sim N_p(\boldsymbol{m}, \Sigma)$ then

$$E\left(e^{i\boldsymbol{t}\cdot\boldsymbol{X}}\right) = e^{i\boldsymbol{t}\cdot\boldsymbol{m}} e^{-\frac{1}{2}\boldsymbol{t}^*\Sigma\boldsymbol{t}} \tag{28.2}$$

If a is a constant and $\boldsymbol{X} \sim N_p(\boldsymbol{m}, \Sigma)$, then $a\boldsymbol{X} \sim N_p\left(a\boldsymbol{m}, a^2\Sigma\right)$.

Proof: Consider $E\left(e^{i\boldsymbol{t}\cdot\boldsymbol{X}}\right)$ for $\boldsymbol{X}\sim N_p(\boldsymbol{m},\Sigma)$.

$$E\left(e^{i\boldsymbol{t}\cdot\boldsymbol{X}}\right)\equiv\frac{1}{(2\pi)^{p/2}\left(\det\Sigma\right)^{1/2}}\int_{\mathbb{R}^p}e^{i\boldsymbol{t}\cdot\boldsymbol{x}}e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^*\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{m})}dx.$$

Let R be an orthogonal transformation such that

$$R\Sigma R^*=D=\operatorname{diag}\left(\sigma_1^2,\cdots,\sigma_p^2\right).$$

Let $R(\boldsymbol{x}-\boldsymbol{m})=\boldsymbol{y}$. Then

$$E\left(e^{i\boldsymbol{t}\cdot\boldsymbol{X}}\right)=\frac{1}{(2\pi)^{p/2}\prod_{i=1}^p\sigma_i}\int_{\mathbb{R}^p}e^{i\boldsymbol{t}\cdot(R^*\boldsymbol{y}+\boldsymbol{m})}e^{-\frac{1}{2}\boldsymbol{y}^*D^{-1}\boldsymbol{y}}dy.$$

Therefore

$$E\left(e^{i\boldsymbol{t}\cdot\boldsymbol{X}}\right)=\frac{1}{(2\pi)^{p/2}\prod_{i=1}^p\sigma_i}\int_{\mathbb{R}^p}e^{i\boldsymbol{s}\cdot(\boldsymbol{y}+R\boldsymbol{m})}e^{-\frac{1}{2}\boldsymbol{y}^*D^{-1}\boldsymbol{y}}dy$$

where $\boldsymbol{s}=R\boldsymbol{t}$. This equals

$$e^{i\boldsymbol{t}\cdot\boldsymbol{m}}\prod_{i=1}^p\left(\int_{\mathbb{R}}e^{is_iy_i}e^{-\frac{1}{2\sigma_i^2}y_i^2}dy_i\right)\frac{1}{\sqrt{2\pi}\sigma_i}$$

$$=e^{i\boldsymbol{t}\cdot\boldsymbol{m}}\prod_{i=1}^p\left(\int_{\mathbb{R}}e^{is_i\sigma_iu}e^{-\frac{1}{2}u^2}du\right)\frac{1}{\sqrt{2\pi}}$$

$$=e^{i\boldsymbol{t}\cdot\boldsymbol{m}}\prod_{i=1}^pe^{-\frac{1}{2}s_i^2\sigma_i^2}\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}e^{-\frac{1}{2}(u-is_i\sigma_i)^2}du$$

By Lemma 28.0.1, this equals $e^{i\boldsymbol{t}\cdot\boldsymbol{m}}e^{-\frac{1}{2}\sum_{i=1}^p s_i^2\sigma_i^2}=e^{i\boldsymbol{t}\cdot\boldsymbol{m}}e^{-\frac{1}{2}\boldsymbol{t}^*\Sigma\boldsymbol{t}}$. This proves 28.2.

Since $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent, $e^{i\boldsymbol{t}\cdot\boldsymbol{X}_1}$ and $e^{i\boldsymbol{t}\cdot\boldsymbol{X}_2}$ are also independent. Hence

$$E\left(e^{i\boldsymbol{t}\cdot\boldsymbol{X}_1+\boldsymbol{X}_2}\right)=E\left(e^{i\boldsymbol{t}\cdot\boldsymbol{X}_1}\right)E\left(e^{i\boldsymbol{t}\cdot\boldsymbol{X}_2}\right).$$

Thus,

$$\begin{aligned}E\left(e^{i\boldsymbol{t}\cdot\boldsymbol{X}_1+\boldsymbol{X}_2}\right)&=E\left(e^{i\boldsymbol{t}\cdot\boldsymbol{X}_1}\right)E\left(e^{i\boldsymbol{t}\cdot\boldsymbol{X}_2}\right)=e^{i\boldsymbol{t}\cdot\boldsymbol{m}_1}e^{-\frac{1}{2}\boldsymbol{t}^*\Sigma_1\boldsymbol{t}}e^{i\boldsymbol{t}\cdot\boldsymbol{m}_2}e^{-\frac{1}{2}\boldsymbol{t}^*\Sigma_2\boldsymbol{t}}\\&=e^{i\boldsymbol{t}\cdot(\boldsymbol{m}_1+\boldsymbol{m}_2)}e^{-\frac{1}{2}\boldsymbol{t}^*(\Sigma_1+\Sigma_2)\boldsymbol{t}},\end{aligned}$$

which, as shown above is the characteristic function of a random vector distributed as $N_p(\boldsymbol{m}_1+\boldsymbol{m}_2,\Sigma_1+\Sigma_2)$. Now it follows that $\boldsymbol{X}_1+\boldsymbol{X}_2\sim N_p(\boldsymbol{m}_1+\boldsymbol{m}_2,\Sigma_1+\Sigma_2)$ by Theorem 27.1.4. This proves 28.1.

The assertion about $-\boldsymbol{X}$ is also easy to see because

$$\begin{aligned}E\left(e^{i\boldsymbol{t}\cdot(-\boldsymbol{X})}\right)&=E\left(e^{i(-\boldsymbol{t})\cdot\boldsymbol{X}}\right)\\&=\frac{1}{(2\pi)^{p/2}\left(\det\Sigma\right)^{1/2}}\int_{\mathbb{R}^p}e^{i(-\boldsymbol{t})\cdot\boldsymbol{x}}e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^*\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{m})}dx\\&=\frac{1}{(2\pi)^{p/2}\left(\det\Sigma\right)^{1/2}}\int_{\mathbb{R}^p}e^{i\boldsymbol{t}\cdot\boldsymbol{x}}e^{-\frac{1}{2}(\boldsymbol{x}+\boldsymbol{m})^*\Sigma^{-1}(\boldsymbol{x}+\boldsymbol{m})}dx\end{aligned}$$

which is the characteristic function of a random variable which is $N(-\boldsymbol{m},\Sigma)$. Theorem 27.1.4 again implies $-\boldsymbol{X} \sim N(-\boldsymbol{m},\Sigma)$. Finally consider the last claim. You apply what is known about $\boldsymbol{X}$ with $\boldsymbol{t}$ replaced with $a\boldsymbol{t}$ and then massage things. This gives the characteristic function for $a\boldsymbol{X}$ is given by

$$E\left(\exp\left(i\,\boldsymbol{t}\cdot a\boldsymbol{X}\right)\right) = \exp\left(i\,\boldsymbol{t}\cdot a\boldsymbol{m}\right)\exp\left(-\frac{1}{2}\boldsymbol{t}^{*}\Sigma a^{2}\boldsymbol{t}\right)$$

which is the characteristic function of a normal random vector having mean $a\boldsymbol{m}$ and covariance $a^2\Sigma$. ■

28.2 Linear Combinations

Following [43] a random vector has a generalized normal distribution if its characteristic function is given as $e^{i\boldsymbol{t}\cdot\boldsymbol{m}}e^{-\frac{1}{2}\boldsymbol{t}^{*}\Sigma\boldsymbol{t}}$ where Σ is symmetric and has nonnegative eigenvalues. For a random real valued variable, $\boldsymbol{m}$ is scalar and so is Σ so the characteristic function of such a generalized normally distributed random variable is $e^{it\mu}e^{-\frac{1}{2}t^2\sigma^2}$. These generalized normal distributions do not require Σ to be invertible, only that the eigenvalues be nonnegative. In one dimension this would correspond the characteristic function of a dirac measure having point mass 1 at μ. In higher dimensions, it could be a mixture of such things with more familiar things. I won't try very hard to distinguish between generalized normal distributions and normal distributions in which the covariance matrix has all positive eigenvalues. These generalized normal distributions are discussed more a little later.

Here are some other interesting results about normal distributions found in [43]. The next theorem has to do with the question whether a random vector is normally distributed in the above generalized sense.

Theorem 28.2.1 *Let $\boldsymbol{X} = (X_1,\cdots,X_p)$ where each X_i is a real valued random variable. Then $\boldsymbol{X}$ is normally distributed in the above generalized sense if and only if every linear combination, $\sum_{j=1}^{p} a_i X_i$ is normally distributed. In this case the mean of $\boldsymbol{X}$ is*

$$\boldsymbol{m} = \left(E\left(X_1\right),\cdots,E\left(X_p\right)\right)$$

and the covariance matrix for $\boldsymbol{X}$ is

$$\Sigma_{jk} = E\left(\left(X_j - m_j\right)\left(X_k - m_k\right)^{*}\right).$$

Proof: Suppose first $\boldsymbol{X}$ is normally distributed. Then its characteristic function is of the form

$$\phi_{\boldsymbol{X}}\left(\boldsymbol{t}\right) = E\left(e^{i\boldsymbol{t}\cdot\boldsymbol{X}}\right) = e^{i\boldsymbol{t}\cdot\boldsymbol{m}}e^{-\frac{1}{2}\boldsymbol{t}^{*}\Sigma\boldsymbol{t}}.$$

Then letting $\boldsymbol{a} = (a_1,\cdots,a_p)$

$$E\left(e^{it\sum_{j=1}^{p} a_i X_i}\right) = E\left(e^{it\boldsymbol{a}\cdot\boldsymbol{X}}\right) = e^{it\boldsymbol{a}\cdot\boldsymbol{m}}e^{-\frac{1}{2}\boldsymbol{a}^{*}\Sigma\boldsymbol{a}t^2}$$

which is the characteristic function of a normally distributed random variable with mean $\boldsymbol{a}\cdot\boldsymbol{m}$ and variance $\sigma^2 = \boldsymbol{a}^{*}\Sigma\boldsymbol{a}$**.** This proves half of the theorem. If $\boldsymbol{X}$ is normally distributed, then every linear combination is normally distributed.

Conversely, suppose every linear combination is normally distributed. Next suppose $\sum_{j=1}^{p} a_j X_j = \boldsymbol{a}\cdot\boldsymbol{X}$ is normally distributed with mean μ and variance σ^2 so that its characteristic function is given as $e^{it\mu}e^{-\frac{1}{2}t^2\sigma^2}$. I will now relate μ and σ^2 to various quantities involving the X_j. Letting $m_j = E(X_j)$, $\boldsymbol{m} = (m_1,\cdots,m_p)^*$

$$\begin{aligned}\mu &= \sum_{j=1}^{p} a_j E(X_j) = \sum_{j=1}^{p} a_j m_j,\ \sigma^2 = E\left(\left(\sum_{j=1}^{p} a_j X_j - \sum_{j=1}^{p} a_j m_j\right)^2\right)\\ &= E\left(\left(\sum_{j=1}^{p} a_j (X_j - m_j)\right)^2\right) = \sum_{j,k} a_j a_k E((X_j - m_j)(X_k - m_k))\end{aligned}$$

It follows the mean of the random variable, $\boldsymbol{a}\cdot\boldsymbol{X}$ is $\mu = \sum_j a_j m_j = \boldsymbol{a}\cdot\boldsymbol{m}$ and its variance is

$$\sigma^2 = \boldsymbol{a}^* E\left((\boldsymbol{X}-\boldsymbol{m})(\boldsymbol{X}-\boldsymbol{m})^*\right)\boldsymbol{a}$$

Therefore, $E\left(e^{it\boldsymbol{a}\cdot\boldsymbol{X}}\right) = e^{it\mu}e^{-\frac{1}{2}t^2\sigma^2} = e^{it\boldsymbol{a}\cdot\boldsymbol{m}}e^{-\frac{1}{2}t^2\boldsymbol{a}^*E\left((\boldsymbol{X}-\boldsymbol{m})(\boldsymbol{X}-\boldsymbol{m})^*\right)\boldsymbol{a}}$. Letting $\boldsymbol{s} = t\boldsymbol{a}$ this shows

$$E\left(e^{i\boldsymbol{s}\cdot\boldsymbol{X}}\right) = e^{i\boldsymbol{s}\cdot\boldsymbol{m}}e^{-\frac{1}{2}\boldsymbol{s}^*E\left((\boldsymbol{X}-\boldsymbol{m})(\boldsymbol{X}-\boldsymbol{m})^*\right)\boldsymbol{s}} = e^{i\boldsymbol{s}\cdot\boldsymbol{m}}e^{-\frac{1}{2}\boldsymbol{s}^*\Sigma\boldsymbol{s}}$$

which is the characteristic function of a normally distributed random variable with $\boldsymbol{m}$ given above and Σ given by

$$\Sigma_{jk} = E((X_j - m_j)(X_k - m_k)).$$

By assumption, $\boldsymbol{a}$ is completely arbitrary and so it follows that $\boldsymbol{s}$ is also. Hence, $\boldsymbol{X}$ is normally distributed as claimed. ■

Corollary 28.2.2 *Let $\boldsymbol{X} = (X_1,\cdots,X_p)$, $\boldsymbol{Y} = (Y_1,\cdots,Y_p)$ where each X_i, Y_i is a real valued random variable. Suppose also that for every $\boldsymbol{a}\in\mathbb{R}^p$, $\boldsymbol{a}\cdot\boldsymbol{X}$ and $\boldsymbol{a}\cdot\boldsymbol{Y}$ are both normally distributed with the same mean and variance. Then $\boldsymbol{X}$ and $\boldsymbol{Y}$ are both multivariate normal random vectors with the same mean and variance.*

Proof: In the Proof of Theorem 28.2.1 the proof implies that the characteristic functions of $\boldsymbol{a}\cdot\boldsymbol{X}$ and $\boldsymbol{a}\cdot\boldsymbol{Y}$ are both of the form $e^{itm}e^{-\frac{1}{2}\sigma^2t^2}$. Then as in the proof of that theorem, it must be the case that $m = \sum_{j=1}^{p} a_j m_j$ where $E(X_i) = m_i = E(Y_i)$ and

$$\sigma^2 = \boldsymbol{a}^* E\left((\boldsymbol{X}-\boldsymbol{m})(\boldsymbol{X}-\boldsymbol{m})^*\right)\boldsymbol{a} = \boldsymbol{a}^* E\left((\boldsymbol{Y}-\boldsymbol{m})(\boldsymbol{Y}-\boldsymbol{m})^*\right)\boldsymbol{a}$$

and this last equation must hold for every $\boldsymbol{a}$. Therefore,

$$E\left((\boldsymbol{X}-\boldsymbol{m})(\boldsymbol{X}-\boldsymbol{m})^*\right) = E\left((\boldsymbol{Y}-\boldsymbol{m})(\boldsymbol{Y}-\boldsymbol{m})^*\right) \equiv \Sigma$$

and so the characteristic function of both $\boldsymbol{X}$ and $\boldsymbol{Y}$ is $e^{i\boldsymbol{s}\cdot\boldsymbol{m}}e^{-\frac{1}{2}\boldsymbol{s}^*\Sigma\boldsymbol{s}}$ as in the proof of Theorem 28.2.1. ■

Theorem 28.2.3 *Suppose $\boldsymbol{X} = (X_1,\cdots,X_p)$ is normally distributed with mean $\boldsymbol{m}$ and covariance Σ. Then if X_1 is uncorrelated with any of the X_i, meaning*

$$E((X_1 - m_1)(X_j - m_j)) = 0 \text{ for } j > 1,$$

then X_1 and $(X_2,\cdots,X_p)$ are both normally distributed and the two random vectors are independent. Here $m_j \equiv E(X_j)$. More generally, if the covariance matrix is a diagonal matrix, the random variables, $\{X_1,\cdots,X_p\}$ are independent.

Proof: From Theorem 28.1.2 $\Sigma = E\left((\boldsymbol{X}-\boldsymbol{m})(\boldsymbol{X}-\boldsymbol{m})^*\right)$. Then by assumption,

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \mathbf{0} \\ \mathbf{0} & \Sigma_{p-1} \end{pmatrix}. \tag{28.3}$$

I need to verify that if $E \in \sigma(X_1)$ and $F \in \sigma(X_2,\cdots,X_p)$, then $P(E\cap F) = P(E)P(F)$.

Let $E = X_1^{-1}(A)$ and $F = (X_2,\cdots,X_p)^{-1}(B)$ where A and B are Borel sets in $\mathbb{R}$ and $\mathbb{R}^{p-1}$ respectively. Thus I need to verify that

$$P([(X_1,(X_2,\cdots,X_p))\in(A,B)]) =$$

$$\mu_{(X_1,(X_2,\cdots,X_p))}(A\times B) = \mu_{X_1}(A)\,\mu_{(X_2,\cdots,X_p)}(B). \tag{28.4}$$

Using 28.3, Fubini's theorem, and definitions,

$$\mu_{(X_1,(X_2,\cdots,X_p))}(A\times B) =$$

$$\begin{aligned}
&\int_{\mathbb{R}^p} \mathcal{X}_{A\times B}(\boldsymbol{x}) \frac{1}{(2\pi)^{p/2}\det(\Sigma)^{1/2}} e^{\frac{-1}{2}(\boldsymbol{x}-\boldsymbol{m})^*\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{m})}dx \\
&= \int_{\mathbb{R}} \mathcal{X}_A(x_1)\int_{\mathbb{R}^{p-1}} \mathcal{X}_B(X_2,\cdots,X_p)\cdot \\
&\quad \frac{1}{(2\pi)^{(p-1)/2}\sqrt{2\pi}\left(\sigma_1^2\right)^{1/2}\det(\Sigma_{p-1})^{1/2}} e^{\frac{-(x_1-m_1)^2}{2\sigma_1^2}}\cdot \\
&\quad e^{\frac{-1}{2}\left(\boldsymbol{x}'-\boldsymbol{m}'\right)^*\Sigma_{p-1}^{-1}\left(\boldsymbol{x}'-\boldsymbol{m}'\right)}dx'dx_1
\end{aligned}$$

where $\boldsymbol{x}' = (x_2,\cdots,x_p)$ and $\boldsymbol{m}' = (m_2,\cdots,m_p)$. Now this equals

$$\begin{aligned}
&\int_{\mathbb{R}} \mathcal{X}_A(x_1)\frac{1}{\sqrt{2\pi\sigma_1^2}} e^{\frac{-(x_1-m_1)^2}{2\sigma_1^2}}\cdot \\
&\int_B \frac{1}{(2\pi)^{(p-1)/2}\det(\Sigma_{p-1})^{1/2}} e^{\frac{-1}{2}\left(\boldsymbol{x}'-\boldsymbol{m}'\right)^*\Sigma_{p-1}^{-1}\left(\boldsymbol{x}'-\boldsymbol{m}'\right)}dx'dx.
\end{aligned} \tag{28.5}$$

In case $B = \mathbb{R}^{p-1}$, the inside integral equals 1 and

$$\mu_{X_1}(A) = \mu_{(X_1,(X_2,\cdots,X_p))}\left(A\times\mathbb{R}^{p-1}\right) = \int_{\mathbb{R}} \mathcal{X}_A(x_1)\frac{1}{\sqrt{2\pi\sigma_1^2}} e^{\frac{-(x_1-m_1)^2}{2\sigma_1^2}}dx_1$$

which shows X_1 is normally distributed as claimed. Similarly, letting $A = \mathbb{R}$,

$$\begin{aligned}
&\mu_{(X_2,\cdots,X_p)}(B) = \mu_{(X_1,(X_2,\cdots,X_p))}(\mathbb{R}\times B) \\
&= \int_B \frac{1}{(2\pi)^{(p-1)/2}\det(\Sigma_{p-1})^{1/2}} e^{\frac{-1}{2}\left(\boldsymbol{x}'-\boldsymbol{m}'\right)^*\Sigma_{p-1}^{-1}\left(\boldsymbol{x}'-\boldsymbol{m}'\right)}dx'
\end{aligned}$$

and $(X_2,\cdots,X_p)$ is also normally distributed with mean $\boldsymbol{m}'$ and covariance Σ_{p-1}. Now from 28.5, 28.4 follows. In case the covariance matrix is diagonal, the above reasoning extends in an obvious way to prove the random variables, $\{X_1,\cdots,X_p\}$ are independent.

However, another way to prove this is to use Proposition 27.4.1 on Page 773 and consider the characteristic function. Let $E(X_j) = m_j$ and $P = \sum_{j=1}^{p} t_j X_j$. Then since $\boldsymbol{X}$ is normally distributed and the covariance is a diagonal,

$$D \equiv \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_p^2 \end{pmatrix}$$

,

$$\begin{aligned} E\left(e^{iP}\right) &= E\left(e^{i\boldsymbol{t}\cdot\boldsymbol{X}}\right) = e^{i\boldsymbol{t}\cdot\boldsymbol{m}} e^{-\frac{1}{2}\boldsymbol{t}^*\Sigma\boldsymbol{t}} = \exp\left(\sum_{j=1}^{p} it_j m_j - \frac{1}{2}t_j^2\sigma_j^2\right) \\ &= \prod_{j=1}^{p} \exp\left(it_j m_j - \frac{1}{2}t_j^2\sigma_j^2\right) \end{aligned} \tag{28.6}$$

Also,

$$E\left(e^{it_j X_j}\right) = E\left(\exp\left(it_j X_j + \sum_{k\neq j} i0X_k\right)\right) = \exp\left(it_j m_j - \frac{1}{2}t_j^2\sigma_j^2\right)$$

With 28.6, this shows $E\left(e^{iP}\right) = \prod_{j=1}^{p} E\left(e^{it_j X_j}\right)$ which shows by Proposition 27.4.1 that the random variables, $\{X_1, \cdots, X_p\}$ are independent. ■

28.3 Finding Moments

Let X be a random variable with characteristic function

$$\phi_X(t) \equiv E(\exp(itX))$$

Then this can be used to find moments of the random variable assuming they exist. The k^{th} moment is defined as $E\left(X^k\right)$. This can be done by using the dominated convergence theorem to differentiate the characteristic function with respect to t and then plugging in $t = 0$. For example, $\phi_X'(t) = E(iX\exp(itX))$ and now plugging in $t = 0$ you get $iE(X)$. Doing another differentiation you obtain $\phi_X''(t) = E\left(-X^2\exp(itX)\right)$ and plugging in $t = 0$ you get $-E\left(X^2\right)$ and so forth.

An important case is where X is normally distributed with mean 0 and variance σ^2. In this case, as shown above, the characteristic function is $e^{-\frac{1}{2}t^2\sigma^2}$. Also all moments exist when X is normally distributed. So what are these moments? $D_t\left(e^{-\frac{1}{2}t^2\sigma^2}\right) = -t\sigma^2 e^{-\frac{1}{2}t^2\sigma^2}$ and plugging in $t = 0$ you find the mean equals 0 as expected.

$$D_t\left(-t\sigma^2 e^{-\frac{1}{2}t^2\sigma^2}\right) = -\sigma^2 e^{-\frac{1}{2}t^2\sigma^2} + t^2\sigma^4 e^{-\frac{1}{2}t^2\sigma^2}$$

and plugging in $t = 0$ you find the second moment is σ^2. Then do it again.

$$D_t\left(-\sigma^2 e^{-\frac{1}{2}t^2\sigma^2} + t^2\sigma^4 e^{-\frac{1}{2}t^2\sigma^2}\right) = 3\sigma^4 t e^{-\frac{1}{2}t^2\sigma^2} - t^3\sigma^6 e^{-\frac{1}{2}t^2\sigma^2}$$

Then $E\left(X^3\right) = 0$.

$$D_t\left(3\sigma^4 t e^{-\frac{1}{2}t^2\sigma^2} - t^3\sigma^6 e^{-\frac{1}{2}t^2\sigma^2}\right) = 3\sigma^4 e^{-\frac{1}{2}t^2\sigma^2} - 6\sigma^6 t^2 e^{-\frac{1}{2}t^2\sigma^2} + t^4\sigma^8 e^{-\frac{1}{2}t^2\sigma^2}$$

and so $E\left(X^4\right) = 3\sigma^4$. By now you can see the pattern. If you continue this way, you find the odd moments are all 0 and

$$E\left(X^{2m}\right) = C_m\left(\sigma^2\right)^m. \tag{28.7}$$

This is an important observation. In the case of $\boldsymbol{X}$ a random vector, you have $\phi_{\boldsymbol{X}}(\boldsymbol{t}) \equiv E(\exp(i\boldsymbol{t}\cdot\boldsymbol{X}))$ and by taking $\frac{d}{dt_j}$, you can follow the above procedure to obtain $E(X_j)$ and then by using successive differentiations, you can find $E(X_i^n)$ or any polynomial in the X_i assuming the expectations exist.

28.4 Prokhorov and Levy Theorems

Recall one can define the characteristic function of a probability measure μ as $\int_{\mathbb{R}^p} e^{i\boldsymbol{t}\cdot\boldsymbol{x}}d\mu$. In a sense it is more natural. One can also generalize to replace $\mathbb{R}^p$ with E a Banach space in which the dot product $\boldsymbol{t}\cdot\boldsymbol{x}$ is replaced with $t(x)$ where $t \in E'$. However, the main interest here is in $\mathbb{R}^p$.

Definition 28.4.1 *A set of Borel probability measures $\{\mu_n\}_{n=1}^{\infty}$ defined on a Polish space E is called "tight" if for all $\varepsilon > 0$ there exists a compact set, K_ε such that*

$$\mu_n([\boldsymbol{x} \notin K_\varepsilon]) < \varepsilon$$

for all μ_n.

How do you determine in general that a set of probability measures is tight?

Lemma 28.4.2 *Let E be a separable complete metric space and let Λ be a set of Borel probability measures. Then Λ is tight if and only if for every $\varepsilon > 0$ and $r > 0$ there exists a finite collection of balls, $\{B(a_i, r)\}_{i=1}^m$ such that*

$$\mu\left(\cup_{i=1}^m \overline{B(a_i, r)}\right) > 1 - \varepsilon$$

for every $\mu \in \Lambda$.

Proof: If Λ is tight, then there exists a compact set, K_ε such that

$$\mu(K_\varepsilon) > 1 - \varepsilon$$

for all $\mu \in \Lambda$. Then consider the open cover, $\{B(x, r) : x \in K_\varepsilon\}$. Finitely many of these cover K_ε and this yields the above condition.

Now suppose the above condition and let

$$C_n \equiv \cup_{i=1}^{m_n} \overline{B(a_i^n, 1/n)}$$

satisfy $\mu(C_n) > 1 - \varepsilon/2^n$ for all $\mu \in \Lambda$. Then let $K_\varepsilon \equiv \cap_{n=1}^{\infty} C_n$. This set K_ε is a compact set because it is a closed subset of a complete metric space and is therefore complete, and it is also totally bounded by construction. For $\mu \in \Lambda$,

$$\mu\left(K_\varepsilon^C\right) = \mu\left(\cup_n C_n^C\right) \leq \sum_n \mu\left(C_n^C\right) < \sum_n \frac{\varepsilon}{2^n} = \varepsilon$$

Therefore, Λ is tight. ■

In case the Polish space is $\mathbb{R}^p$, the case of most interest, there is a very nice condition interms of characteristic functions which gives "tightness".

Lemma 28.4.3 *If $\{\mu_n\}$ is a sequence of Borel probability measures defined on the Borel sets of $\mathbb{R}^p$ such that*

$$\lim_{n\to\infty} \phi_{\mu_n}(\boldsymbol{t}) = \psi(\boldsymbol{t})$$

for all $\boldsymbol{t}$, where $\psi(\mathbf{0}) = 1$ and ψ is continuous at $\mathbf{0}$, then $\{\mu_n\}_{n=1}^{\infty}$ is tight.

Proof: Let $\boldsymbol{e}_j$ be the j^{th} standard unit basis vector. Letting $\boldsymbol{t} = t\boldsymbol{e}_j$ in the definition and $u > 0$

$$\left|\frac{1}{u}\int_{-u}^{u}\left(1-\phi_{\mu_n}(t\boldsymbol{e}_j)\right)dt\right| = \left|\frac{1}{u}\int_{-u}^{u}\left(1-\int_{\mathbb{R}^p} e^{itx_j}d\mu_n(x)\right)dt\right| \tag{28.8}$$

$$= \left|\frac{1}{u}\int_{-u}^{u}\left(\int_{\mathbb{R}^p}\left(1-e^{itx_j}\right)d\mu_n(x)\right)dt\right| = \left|\int_{\mathbb{R}^p}\frac{1}{u}\int_{-u}^{u}\left(1-e^{itx_j}\right)dtd\mu_n(x)\right|$$

$$= \left|2\int_{\mathbb{R}^p}\left(1-\frac{\sin(ux_j)}{ux_j}\right)d\mu_n(x)\right| \geq 2\int_{\left[|x_j|\geq\frac{2}{u}\right]}\left(1-\frac{1}{|ux_j|}\right)d\mu_n(x)$$

$$\geq 2\int_{\left[|x_j|\geq\frac{2}{u}\right]}\left(1-\frac{1}{u(2/u)}\right)d\mu_n(x) = \int_{\left[|x_j|\geq\frac{2}{u}\right]}1d\mu_n(x) = \mu_n\left(\left[\boldsymbol{x}:|x_j|\geq\frac{2}{u}\right]\right).$$

If $\varepsilon > 0$ is given, there exists $r > 0$ such that if $u \leq r$, $\frac{1}{u}\int_{-u}^{u}(1-\psi(t\boldsymbol{e}_j))\,dt < \varepsilon/p$ for all $j = 1,\cdots,p$ and so, by the dominated convergence theorem, the same is true with ϕ_{μ_n} in place of ψ provided n is large enough, say $n \geq N(r)$. Thus, from 28.8, if $n \geq N(r)$, $\mu_n\left(\left[\boldsymbol{x}:|x_j|>2r\right]\right) < \varepsilon/p$ for all $j \in \{1,\cdots,p\}$. It follows for $n \geq N(r)$,

$$\mu_n([\boldsymbol{x}:\|\boldsymbol{x}\|_\infty > 2r]) < \varepsilon.$$

and so let $K_\varepsilon \equiv [-r,r]^p$. ■

In the case of $\mathbb{R}^p$, and μ_n, μ Borel probability measures, convergence of characteristic functions yields something interesting for $\psi \in \mathscr{G}$ or $\mathfrak{S}$, the Schwartz class.

Lemma 28.4.4 *If $\phi_{\mu_n}(\boldsymbol{t}) \to \phi_\mu(\boldsymbol{t})$ for all $\boldsymbol{t}$, then whenever $\psi \in \mathfrak{S}$, the Schwartz class,*

$$\mu_n(\psi) \equiv \int_{\mathbb{R}^p}\psi(\boldsymbol{y})\,d\mu_n(y) \to \int_{\mathbb{R}^p}\psi(\boldsymbol{y})\,d\mu(y) \equiv \mu(\psi)$$

as $n \to \infty$.

Proof: By definition, $\phi_\mu(\boldsymbol{y}) \equiv \int_{\mathbb{R}^p} e^{i\boldsymbol{y}\cdot\boldsymbol{x}}d\mu(x)$. Also remember the inverse Fourier transform. Letting $\psi \in \mathfrak{S}$, the Schwartz class,

$$\begin{aligned} F^{-1}(\mu)(\psi) &\equiv \mu\left(F^{-1}\psi\right) \equiv \int_{\mathbb{R}^p} F^{-1}\psi d\mu \\ &= \frac{1}{(2\pi)^{p/2}}\int_{\mathbb{R}^p}\int_{\mathbb{R}^p} e^{i\boldsymbol{y}\cdot\boldsymbol{x}}\psi(\boldsymbol{x})\,dxd\mu(y) \end{aligned}$$

$$= \frac{1}{(2\pi)^{p/2}}\int_{\mathbb{R}^p}\psi(\boldsymbol{x})\int_{\mathbb{R}^p} e^{i\boldsymbol{y}\cdot\boldsymbol{x}}d\mu(y)\,dx = \frac{1}{(2\pi)^{p/2}}\int_{\mathbb{R}^p}\psi(\boldsymbol{x})\,\phi_\mu(\boldsymbol{x})\,dx$$

and so, considered as elements of $\mathfrak{S}^*$ or $\mathscr{G}^*$, $F^{-1}(\mu) = \phi_\mu(\cdot)(2\pi)^{-(p/2)} \in L^\infty$. By the dominated convergence theorem

$$\begin{aligned} (2\pi)^{p/2}F^{-1}(\mu_n)(\psi) &\equiv \int_{\mathbb{R}^p}\phi_{\mu_n}(\boldsymbol{t})\,\psi(\boldsymbol{t})\,dt \to \int_{\mathbb{R}^p}\phi_\mu(\boldsymbol{t})\,\psi(\boldsymbol{t})\,dt \\ &= (2\pi)^{p/2}F^{-1}(\mu)(\psi) \end{aligned}$$

whenever $\psi \in \mathfrak{S}$ or $\mathscr{G}$. Thus

$$\begin{aligned}\mu_n(\psi) &= FF^{-1}\mu_n(\psi) \equiv F^{-1}\mu_n(F\psi) \to F^{-1}\mu(F\psi)\\ &\equiv F^{-1}F\mu(\psi) = \mu(\psi). \blacksquare\end{aligned}$$

The set $\mathscr{G}$ of $\mathfrak{S}$ generalizes to ψ any bounded uniformly continuous function.

Lemma 28.4.5 *If $\phi_{\mu_n}(t) \to \phi_\mu(t)$ where $\{\mu_n\}$ and μ are probability measures defined on the Borel sets of $\mathbb{R}^p$, then if ψ is any bounded uniformly continuous function,*

$$\lim_{n\to\infty}\int_{\mathbb{R}^p}\psi d\mu_n = \int_{\mathbb{R}^p}\psi d\mu.$$

Proof: Let $\varepsilon > 0$ be given, let ψ be a bounded function in $C^\infty(\mathbb{R}^p)$. Now let $\eta \in C_c^\infty(Q_r)$ where $Q_r \equiv [-r,r]^p$ satisfy the additional requirement that $\eta = 1$ on $Q_{r/2}$ and $\eta(\boldsymbol{x}) \in [0,1]$ for all $\boldsymbol{x}$. By Lemma 28.4.3 the set, $\{\mu_n\}_{n=1}^\infty$, is tight and so if $\varepsilon > 0$ is given, there exists r sufficiently large such that for all n,

$$\int_{[\boldsymbol{x}\notin Q_{r/2}]} |1-\eta|\,|\psi|\,d\mu_n < \frac{\varepsilon}{3},$$

and since μ is a single measure, the following holds whenever r is large enough.

$$\int_{[\boldsymbol{x}\notin Q_{r/2}]} |1-\eta|\,|\psi|\,d\mu < \frac{\varepsilon}{3}.$$

Thus,

$$\begin{aligned}\left|\int_{\mathbb{R}^p}\psi d\mu_n - \int_{\mathbb{R}^p}\psi d\mu\right| &\le \left|\int_{\mathbb{R}^p}\psi d\mu_n - \int_{\mathbb{R}^p}\psi\eta d\mu_n\right| +\\ \left|\int_{\mathbb{R}^p}\psi\eta d\mu_n - \int_{\mathbb{R}^p}\psi\eta d\mu\right| &+ \left|\int_{\mathbb{R}^p}\psi\eta d\mu - \int_{\mathbb{R}^p}\psi d\mu\right|\\ &\le \frac{2\varepsilon}{3} + \left|\int_{\mathbb{R}^p}\psi\eta d\mu_n - \int_{\mathbb{R}^p}\psi\eta d\mu\right| < \varepsilon\end{aligned}$$

whenever n is large enough by Lemma 28.4.4 because $\psi\eta \in \mathfrak{S}$. This establishes the conclusion of the lemma in the case where ψ is also infinitely differentiable. To consider the general case, let ψ only be uniformly continuous and let $\psi_k = \psi * \phi_k$ where ϕ_k is a mollifier whose support is in $(-(1/k),(1/k))^p$. Then ψ_k converges uniformly to ψ and so the desired conclusion follows for ψ after a routine estimate. $\blacksquare$

Here are some items which are of considerable interest for their own sake.

Theorem 28.4.6 *Let H be a compact metric space. Then there exists a compact subset of $[0,1]$, K and a continuous function, θ which maps K onto H.*

Proof: Without loss of generality, it can be assumed H is an infinite set since otherwise the conclusion is trivial. You could pick finitely many points of $[0,1]$ for K.

Since H is compact, it is totally bounded. Therefore, there exists a 1 net for H $\{h_i\}_{i=1}^{m_1}$. Letting $H_i^1 \equiv \overline{B(h_i,1)}$, it follows H_i^1 is also a compact metric space and so there exists a 1/2 net for each H_i^1, $\left\{h_j^i\right\}_{j=1}^{m_i}$. Then taking the intersection of $\overline{B\left(h_j^i,\frac{1}{2}\right)}$ with H_i^1 to obtain sets

denoted by H_j^2 and continuing this way, one can obtain compact subsets of H, $\left\{H_k^i\right\}$ which satisfies: each H_j^i is contained in some H_k^{i-1}, each H_j^i is compact with diameter less than i^{-1}, each H_j^i is the union of sets of the form H_k^{i+1} which are contained in it. Denoting by $\left\{H_j^i\right\}_{j=1}^{m_i}$ those sets corresponding to a superscript of i, it can also be assumed $m_i < m_{i+1}$. If this is not so, simply add in another point to the i^{-1} net. Now let $\left\{I_j^i\right\}_{j=1}^{m_i}$ be disjoint closed intervals in $[0,1]$ each of length no longer than 2^{-m_i} which have the property that I_j^i is contained in I_k^{i-1} for some k. Letting $K_i \equiv \cup_{j=1}^{m_i} I_j^i$, it follows K_i is a sequence of nested compact sets. Let $K = \cap_{i=1}^{\infty} K_i$. Then each $x \in K$ is the intersection of a unique sequence of these closed intervals, $\left\{I_{j_k}^k\right\}_{k=1}^{\infty}$. Define $\theta x \equiv \cap_{k=1}^{\infty} H_{j_k}^k$. Since the diameters of the H_j^i converge to 0 as $i \to \infty$, this function is well defined. It is continuous because if $x_n \to x$, then ultimately x_n and x are both in $I_{j_k}^k$, the k^{th} closed interval in the sequence whose intersection is x. Hence,

$$d(\theta x_n, \theta x) \leq \text{diameter}(H_{j_k}^k) \leq 1/k.$$

To see the map is onto, let $h \in H$. Then from the construction, there exists a sequence $\left\{H_{j_k}^k\right\}_{k=1}^{\infty}$ of the above sets whose intersection equals h. Then $\theta\left(\cap_{i=1}^{\infty} I_{j_k}^k\right) = h$. ■

Note θ is probably not one to one.

As an important corollary, it follows that the continuous functions defined on any compact metric space is separable.

Corollary 28.4.7 *Let H be a compact metric space and let $C(H)$ denote the continuous functions defined on H with the usual norm,*

$$||f||_{\infty} \equiv \max\{|f(x)| : x \in H\}$$

Then $C(H)$ is separable.

Proof: The proof is by contradiction. Suppose $C(H)$ is not separable. Let $\mathscr{H}_k$ denote a maximal collection of functions of $C(H)$ with the property that if $f, g \in \mathscr{H}_k$, then $||f-g||_{\infty} \geq 1/k$. The existence of such a maximal collection of functions is a consequence of a simple use of the Hausdorff maximality theorem. Then $\cup_{k=1}^{\infty} \mathscr{H}_k$ is dense. Therefore, it cannot be countable by the assumption that $C(H)$ is not separable. It follows that for some k, $\mathscr{H}_k$ is uncountable. Now by Theorem 28.4.6 there exists a continuous function θ defined on a compact subset K of $[0,1]$ which maps K onto H. Now consider the functions defined on K

$$\mathscr{G}_k \equiv \{f \circ \theta : f \in \mathscr{H}_k\}.$$

Then $\mathscr{G}_k$ is an uncountable set of continuous functions defined on K with the property that the distance between any two of them is at least as large as $1/k$. This contradicts separability of $C(K)$ which follows from the Weierstrass approximation theorem in which the separable countable set of functions is the restrictions of polynomials that involve only rational coefficients. ■

The next theorem gives the existence of a measure based on an assumption that a set of measures is tight. It is a sort of sequential compactness result. It is Prokhorov's theorem about a tight set of measures. Recall that Λ is tight means that for every $\varepsilon > 0$ there exists K compact such that $\mu\left(K^C\right) < \varepsilon$ for all $\mu \in \Lambda$.

Theorem 28.4.8 *Let $\Lambda = \{\mu_n\}_{n=1}^{\infty}$ be a sequence of probability measures defined on the Borel sets of E a Polish space. If Λ is tight then there exists a probability measure λ and a subsequence of $\{\mu_n\}_{n=1}^{\infty}$, still denoted by $\{\mu_n\}_{n=1}^{\infty}$ such that whenever ϕ is a continuous bounded complex valued function defined on E,*

$$\lim_{n\to\infty} \int \phi d\mu_n = \int \phi d\lambda.$$

Conversely, if μ_n converges weakly to λ, then $\{\mu_n\}$ is tight.

Proof: By tightness, there exists an increasing sequence of compact sets, $\{K_n\}$ such that $\mu(K_n) > 1 - \frac{1}{n}$ for all $\mu \in \Lambda$. Now letting $\mu \in \Lambda$ and $\phi \in C(K_n)$ such that $\|\phi\|_\infty \leq 1$, it follows

$$\left| \int_{K_n} \phi d\mu \right| \leq \mu(K_n) \leq 1$$

and so the restrictions of the measures of Λ to K_n are contained in the unit ball of $C(K_n)'$. Recall from the Riesz representation theorem, the dual space of $C(K_n)$ is a space of complex Borel measures. Theorem 21.5.5 on Page 581 implies the unit ball of $C(K_n)'$ is weak $*$ sequentially compact. This follows from the observation that $C(K_n)$ is separable which follows easily from the Weierstrass approximation theorem. Recall from the Riesz representation theorem, the dual space of $C(K_n)$ is a space of complex Borel measures. Theorem 21.5.5 on Page 581 implies the unit ball of $C(K_n)'$ is weak $*$ sequentially compact. This follows from the observation that $C(K_n)$ is separable which is proved in Corollary 28.4.7 and leads to the fact that the unit ball in $C(K_n)'$ is actually metrizable by Theorem 21.5.5 on Page 581.

Thus the unit ball in $C(K_n)'$ is actually metrizable by Theorem 21.5.5 on Page 581. Therefore, there exists a subsequence of Λ, $\{\mu_{1k}\}$ such that their restrictions to K_1 converge weak $*$ to a measure, $\lambda_1 \in C(K_1)'$. That is, for every $\phi \in C(K_1)$,

$$\lim_{k\to\infty} \int_{K_1} \phi d\mu_{1k} = \int_{K_1} \phi d\lambda_1$$

By the same reasoning, there exists a further subsequence $\{\mu_{2k}\}$ such that the restrictions of these measures to K_2 converge weak $*$ to a measure $\lambda_2 \in C(K_2)'$ etc. Continuing this way,

$$\begin{array}{l} \mu_{11}, \mu_{12}, \mu_{13}, \cdots \to \text{ Weak} * \text{ in } C(K_1)' \\ \mu_{21}, \mu_{22}, \mu_{23}, \cdots \to \text{ Weak} * \text{ in } C(K_2)' \\ \mu_{31}, \mu_{32}, \mu_{33}, \cdots \to \text{ Weak} * \text{ in } C(K_3)' \\ \vdots \end{array}$$

Here the j^{th} sequence is a subsequence of the $(j-1)^{th}$. Let λ_n denote the measure in $C(K_n)'$ to which the sequence $\{\mu_{nk}\}_{k=1}^{\infty}$ converges weak $*$. Let $\{\mu_n\} \equiv \{\mu_{nn}\}$, the diagonal sequence. Thus this sequence is ultimately a subsequence of every one of the above sequences and so μ_n converges weak $*$ in $C(K_m)'$ to λ_m for each m.

Claim: For $p > n$, the restriction of λ_p to the Borel sets of K_n equals λ_n.

Proof of claim: Let H be a compact subset of K_n. Then there are sets, V_l open in K_n which are decreasing and whose intersection equals H. This follows because this is a metric

space. Then let $H \prec \phi_l \prec V_l$. It follows

$$\begin{aligned}\lambda_n(V_l) &\geq \int_{K_n} \phi_l d\lambda_n = \lim_{k\to\infty} \int_{K_n} \phi_l d\mu_k \\ &= \lim_{k\to\infty} \int_{K_p} \phi_l d\mu_k = \int_{K_p} \phi_l d\lambda_p \geq \lambda_p(H).\end{aligned}$$

Now considering the ends of this inequality, let $l \to \infty$ and pass to the limit to conclude $\lambda_n(H) \geq \lambda_p(H)$. Similarly,

$$\begin{aligned}\lambda_n(H) &\leq \int_{K_n} \phi_l d\lambda_n = \lim_{k\to\infty} \int_{K_n} \phi_l d\mu_k \\ &= \lim_{k\to\infty} \int_{K_p} \phi_l d\mu_k = \int_{K_p} \phi_l d\lambda_p \leq \lambda_p(V_l).\end{aligned}$$

Then passing to the limit as $l \to \infty$, it follows $\lambda_n(H) \leq \lambda_p(H)$. Thus the restriction of $\lambda_p, \lambda_p|_{K_n}$ to the compact sets of K_n equals λ_n. Then by inner regularity it follows the two measures, $\lambda_p|_{K_n}$, and λ_n are equal on all Borel sets of K_n. Recall that for finite measures on the Borel sets of separable metric spaces, regularity is obtained for free.

It is fairly routine to exploit regularity of the measures to verify that $\lambda_m(F) \geq 0$ for all F a Borel subset of K_m. (Whenever $\phi \geq 0, \int_{K_m} \phi d\lambda_m \geq 0$ because $\int_{K_m} \phi d\mu_k \geq 0$. Now you can approximate $\mathscr{X}_F$ with a suitable nonnegative ϕ using regularity of the measure.) Also, letting $\phi \equiv 1$,

$$1 \geq \lambda_m(K_m) \geq 1 - \frac{1}{m}. \tag{28.9}$$

Define for F a Borel set,

$$\lambda(F) \equiv \lim_{n\to\infty} \lambda_n(F \cap K_n).$$

The limit exists because the sequence on the right is increasing due to the above observation that $\lambda_n = \lambda_m$ on the Borel subsets of K_m whenever $n > m$. Thus for $n > m$

$$\lambda_n(F \cap K_n) \geq \lambda_n(F \cap K_m) = \lambda_m(F \cap K_m).$$

Now let $\{F_k\}$ be a sequence of disjoint Borel sets. Then

$$\begin{aligned}\lambda(\cup_{k=1}^\infty F_k) &\equiv \lim_{n\to\infty} \lambda_n(\cup_{k=1}^\infty F_k \cap K_n) = \lim_{n\to\infty} \lambda_n(\cup_{k=1}^\infty (F_k \cap K_n)) \\ &= \lim_{n\to\infty} \sum_{k=1}^\infty \lambda_n(F_k \cap K_n) = \sum_{k=1}^\infty \lambda(F_k)\end{aligned}$$

the last equation holding by the monotone convergence theorem.

It remains to verify $\lim_{k\to\infty} \int \phi d\mu_k = \int \phi d\lambda$ for every ϕ bounded and continuous. This is where tightness is used again. Suppose $\|\phi\|_\infty < M$. Then as noted above, $\lambda_n(K_n) = \lambda(K_n)$ because for $p > n, \lambda_p(K_n) = \lambda_n(K_n)$ and so letting $p \to \infty$, the above is obtained. Also, from 28.9,

$$\begin{aligned}\lambda(K_n^C) &= \lim_{p\to\infty} \lambda_p(K_n^C \cap K_p) \leq \limsup_{p\to\infty} (\lambda_p(K_p) - \lambda_p(K_n)) \\ &\leq \limsup_{p\to\infty} (\lambda_p(K_p) - \lambda_n(K_n)) \leq \limsup_{p\to\infty} \left(1 - \left(1 - \frac{1}{n}\right)\right) = \frac{1}{n}\end{aligned}$$

Consequently,

$$\begin{aligned}\left|\int \phi d\mu_k - \int \phi d\lambda\right| &\leq \left|\int_{K_n^C} \phi d\mu_k + \int_{K_n} \phi d\mu_k - \left(\int_{K_n} \phi d\lambda + \int_{K_n^C} \phi d\lambda\right)\right| \\ &\leq \left|\int_{K_n} \phi d\mu_k - \int_{K_n} \phi d\lambda_n\right| + \left|\int_{K_n^C} \phi d\mu_k - \int_{K_n^C} \phi d\lambda\right| \\ &\leq \left|\int_{K_n} \phi d\mu_k - \int_{K_n} \phi d\lambda_n\right| + \left|\int_{K_n^C} \phi d\mu_k\right| + \left|\int_{K_n^C} \phi d\lambda\right| \\ &\leq \left|\int_{K_n} \phi d\mu_k - \int_{K_n} \phi d\lambda_n\right| + \frac{M}{n} + \frac{M}{n}\end{aligned}$$

First let n be so large that $2M/n < \varepsilon/2$ and then pick k large enough that the above expression is less than ε.

Now suppose μ_n converges to λ weakly. Then for ε there is a compact set such that $\lambda(K) > 1-\varepsilon/2$. This is true because of Lemma 9.4.5 on Page 251 which says that finite measures on a Polish space are inner regular. Then let ψ be a continuous function with values in $[0,1]$ which equals 1 on K and is 0 off a compact set $\hat{K} \supseteq K$. Then $\int \psi d\lambda > 1-\varepsilon/2$ and also, there exists N such that for all $n \geq N$, $\int \psi d\mu_n > 1-\varepsilon/2$. Thus $n \geq N$ implies $\mu_n\left(\hat{K}\right) > 1-\varepsilon/2$. Therefore, enlarging $\hat{K}$ finitely many times, one obtains $\tilde{K} \supseteq K$ such that for all μ_n and $\lambda, \lambda\left(\tilde{K}\right), \mu_n\left(\tilde{K}\right) > 1-\varepsilon/2$. Thus $\mu_n\left(\tilde{K}^C\right) \leq \varepsilon/2 < \varepsilon$ for all n and so $\{\mu_n\}$ is tight as claimed. ■

Definition 28.4.9 *Let $\mu, \{\mu_n\}$ be probability measures defined on the Borel sets of $\mathbb{R}^p$ and let the sequence of probability measures, $\{\mu_n\}$ satisfy*

$$\lim_{n\to\infty} \int \phi d\mu_n = \int \phi d\mu.$$

for every ϕ a bounded continuous function. Then μ_n is said to converge weakly to μ.

With the above, it is possible to prove the following amazing theorem of Levy.

Theorem 28.4.10 *Suppose $\{\mu_n\}$ is a sequence of probability measures defined on the Borel sets of $\mathbb{R}^p$ and let $\left\{\phi_{\mu_n}\right\}$ denote the corresponding sequence of characteristic functions. If there exists ψ which is continuous at $\mathbf{0}$, $\psi(\mathbf{0}) = 1$, and for all $\boldsymbol{t}$,*

$$\phi_{\mu_n}(\boldsymbol{t}) \to \psi(\boldsymbol{t}),$$

then there exists a probability measure λ defined on the Borel sets of $\mathbb{R}^p$ and

$$\phi_\lambda(\boldsymbol{t}) = \psi(\boldsymbol{t}).$$

That is, ψ is a characteristic function of a probability measure. Also, $\{\mu_n\}$ converges weakly to λ.

Proof: By Lemma 28.4.3 $\{\mu_n\}$ is tight. Therefore, there exists a subsequence $\left\{\mu_{n_k}\right\}$ converging weakly to a probability measure λ which implies that

$$\phi_\lambda(\boldsymbol{t}) \equiv \int e^{i\boldsymbol{t}\cdot\boldsymbol{x}} d\lambda(x) = \lim_{n\to\infty} \int e^{i\boldsymbol{t}\cdot\boldsymbol{x}} d\mu_{n_k}(x) = \lim_{n\to\infty} \phi_{\mu_{n_k}}(\boldsymbol{t}) = \psi(\boldsymbol{t}) \blacksquare$$

Note how it was only necessary to assume $\psi(\mathbf{0}) = 1$ and ψ is continuous at $\mathbf{0}$ in order to conclude that ψ is a characteristic function. This helps to see why Prokhorov's and Levy's theorems are so amazing. Limits of characteristic functions tend to be characteristic functions. What about random variables?

If you have a probability measure λ on the Borel sets of $\mathbb{R}^p$, is there a random variable $\boldsymbol{X}$ such that $\lambda = \lambda_{\boldsymbol{X}}$? Yes. You could let $\Omega = \mathbb{R}^p$ and $X(\boldsymbol{x}) = \boldsymbol{x}$ and $P(E) \equiv \lambda(E)$ for all E Borel. Then $\lambda_{\boldsymbol{X}}\left(\boldsymbol{X}^{-1}(E)\right) \equiv P(E) \equiv \lambda(E)$ so this is indeed a random variable such that $\lambda = \lambda_{\boldsymbol{X}}$. Thus for a probability measure λ, you can generally get a random variable which has λ as its distribution measure. Later, this is considered more. You might have more than one random variable having λ as its distribution measure.

In this next corollary, it suffices to have the random variables have values in a Banach space. However, I will write $|\boldsymbol{X}|$ rather than $\|X\|$.

Corollary 28.4.11 *In the context of Theorem 28.4.10, suppose μ_n is the distribution measure of the random variable X_n and that $\sup_n E(|\boldsymbol{X}_n|^q) = M_q < \infty$ for all $q \geq 1$ and that μ_n converges weakly to the probability measure μ. Then if μ is the distribution measure for a random variable $\boldsymbol{X}$, then $E(|\boldsymbol{X}|^q) < \infty$ for all $q \geq 1$.*

Proof:

$$E(|\boldsymbol{X}|^q) = \int_0^\infty P(|\boldsymbol{X}|^q > \alpha)\, d\alpha = \int_0^\infty \mu([|\boldsymbol{x}|^q > \alpha])\, d\alpha$$
$$\leq \int_0^\infty \mu([|\boldsymbol{x}|^q > \alpha])\, d\alpha \leq \int_0^\infty \int_{\mathbb{R}^p} (1 - \psi_\alpha)\, d\mu d\alpha$$

where $\psi_\alpha = 1$ on $B\left(\mathbf{0}, \frac{1}{2}\alpha^{1/q}\right)$ is nonnegative, and is in $C_c\left(B\left(\mathbf{0}, \alpha^{1/q}\right)\right)$. Thus if $|\boldsymbol{x}|^q > \alpha$, then $1 - \psi_\alpha(\boldsymbol{x}) = 1$ which shows the above inequality holds. Also, if $(1 - \psi_\alpha(\boldsymbol{x})) > 0$, then $|\boldsymbol{x}| > \frac{1}{2}\alpha^{1/q}$ and so $|\boldsymbol{x}|^q > \frac{1}{2^q}\alpha$. Since weak convergence holds and $1 - \psi_\alpha$ is a bounded continuous function,

$$\int_{\mathbb{R}^p} (1 - \psi_\alpha)\, d\mu = \lim_{n\to\infty} \int_{\mathbb{R}^p} (1 - \psi_\alpha)\, d\mu_n \tag{28.10}$$

Therefore, from the above and Fatou's lemma,

$$\int_\Omega |\boldsymbol{X}|^q\, dP \leq \int_0^\infty \lim_{n\to\infty} \int_{\mathbb{R}^p} (1 - \psi_\alpha)\, d\mu_n d\alpha$$
$$\leq \liminf_{n\to\infty} \int_0^\infty \mu_n\left(\left[|\boldsymbol{x}|^q > \frac{1}{2^q}\alpha\right]\right) d\alpha$$

Changing the variable,

$$= \liminf_{n\to\infty} 2^q \int_0^\infty \mu_n(|\boldsymbol{x}|^q > \delta)\, d\delta = \liminf_{n\to\infty} 2^q E(|\boldsymbol{X}_n|^q) < \infty \blacksquare$$

Now recall the multivariate normal distribution.

Definition 28.4.12 *A random vector $\boldsymbol{X}$, with values in $\mathbb{R}^p$ has a multivariate normal distribution written as*

$$\boldsymbol{X} \sim N_p(\boldsymbol{m}, \Sigma)$$

if for all Borel $E \subseteq \mathbb{R}^p$, the distribution measure is given by

$$\lambda_{\boldsymbol{X}}(E) = \int_{\mathbb{R}^p} \mathcal{X}_E(\boldsymbol{x}) \frac{1}{(2\pi)^{p/2} \det(\Sigma)^{1/2}} e^{\frac{-1}{2}(\boldsymbol{x}-\boldsymbol{m})^* \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{m})} dx$$

for $\boldsymbol{m}$ a given vector and Σ a given positive definite symmetric matrix. Recall also that the characteristic function of this random variable is

$$E\left(e^{i\boldsymbol{t}\cdot\boldsymbol{X}}\right)=e^{i\boldsymbol{t}\cdot\boldsymbol{m}}e^{-\frac{1}{2}\boldsymbol{t}^*\Sigma\boldsymbol{t}} \tag{28.11}$$

So what if $\det(\Sigma)=0$? Is there a probability measure having characteristic function $e^{i\boldsymbol{t}\cdot\boldsymbol{m}}e^{-\frac{1}{2}\boldsymbol{t}^*\Sigma\boldsymbol{t}}$? Let $\Sigma_n\to\Sigma$ in the Frobenius norm, $\det(\Sigma_n)>0$. That is the ij^{th} components converge and all the eigenvalues are positive. Then from the definition of the characteristic function,

$$\phi_{\lambda_{\boldsymbol{X}_n}}(\boldsymbol{t})=e^{i\boldsymbol{t}\cdot\boldsymbol{m}}e^{-\frac{1}{2}\boldsymbol{t}^*\Sigma_n\boldsymbol{t}}\to\psi(\boldsymbol{t})\equiv e^{i\boldsymbol{t}\cdot\boldsymbol{m}}e^{-\frac{1}{2}\boldsymbol{t}^*\Sigma\boldsymbol{t}}$$

Now clearly $\psi(\mathbf{0})=1$ and ψ is continuous so by Levy's theorem, Theorem 28.4.10, there is a probability measure μ such that $\psi(\boldsymbol{t})=\phi_\mu(\boldsymbol{t})$. As noted above, there is also a random variable $\boldsymbol{X}$ with $\lambda_{\boldsymbol{X}}=\mu$. Consider the moments for $\boldsymbol{X}_n$.

Lemma 28.4.13 *Let $\boldsymbol{X}\sim N(\mathbf{0},\Sigma)$ where Σ is positive definite. Then the moments of $\boldsymbol{X}$ all exist and are dominated by an expression which is continuously dependent on* $\det(\Sigma)$.

Proof: Let $q\geq 1$. $E(|\boldsymbol{X}|^q)=\int_{\mathbb{R}^p}\frac{|\boldsymbol{x}|^q}{(2\pi)^{p/2}\det(\Sigma)^{1/2}}e^{\frac{-1}{2}\boldsymbol{x}^*\Sigma^{-1}\boldsymbol{x}}dx$. Let R be an orthogonal matrix with $\Sigma=R^*DR$ where D is a diagonal matrix having the positive eigenvalues σ_j on the diagonal. Thus $\boldsymbol{x}^*\Sigma^{-1}\boldsymbol{x}=\boldsymbol{x}^*R^*D^{-1}R\boldsymbol{x}$ so let $R\boldsymbol{x}\equiv\boldsymbol{y}$. Changing the variable in the integral and assuming $q=2m$ for m a positive integer,

$$\begin{aligned}E\left(|\boldsymbol{X}|^{2m}\right) &= \int_{\mathbb{R}^p}\frac{\left(\sum_{k=1}^p y_k^2\right)^m}{(2\pi)^{p/2}\prod_{j=1}^p\sigma_j^{1/2}}e^{\frac{-1}{2}\boldsymbol{y}^*D^{-1}\boldsymbol{y}}dy\\ &= 2^p\frac{1}{(2\pi)^{p/2}}\int_0^\infty\cdots\int_0^\infty\frac{\left(\sum_{k=1}^p y_k^2\right)^m}{\prod_{j=1}^p\sigma_j^{1/2}}e^{\frac{-1}{2}\boldsymbol{y}^*D^{-1}\boldsymbol{y}}d\boldsymbol{y}\end{aligned}$$

From convexity of $x\to x^m$

$$\begin{aligned}&= 2^p\frac{1}{(2\pi)^{p/2}}\int_0^\infty\cdots\int_0^\infty\frac{\left(p\sum_{k=1}^p\frac{1}{p}y_k^2\right)^m}{\prod_{j=1}^p\sigma_j^{1/2}}e^{\frac{-1}{2}\boldsymbol{y}^*D^{-1}\boldsymbol{y}}d\boldsymbol{y}\\ &\leq 2^p\frac{p^{m-1}}{(2\pi)^{p/2}}\int_0^\infty\cdots\int_0^\infty\frac{\sum_{k=1}^p y_k^{2m}}{\prod_{j=1}^p\sigma_j^{1/2}}e^{\frac{-1}{2}\sum_{k=1}^p y_k^2\sigma^{-1}}d\boldsymbol{y}\end{aligned}$$

$$\begin{aligned}&= 2^p\frac{p^{m-1}}{(2\pi)^{p/2}}\int_0^\infty\cdots\int_0^\infty\frac{\sum_{k\neq l}y_k^{2m}}{\prod_{j=1}^p\sigma_j^{1/2}}e^{\frac{-1}{2}\sum_{k\neq l}y_k^2\sigma^{-1}}dx_1\cdots\widehat{dx_l}\cdots dx_p\\ &\cdot\int_0^\infty\frac{1}{\sigma_l^{1/2}}y_l^{2m}e^{-\frac{1}{2}y_l^2\sigma_l^{-1}}dy_l\end{aligned}$$

Now letting $u=y_l\sigma_l^{-1/2}, dy_l=\sigma_l^{1/2}du$ and so that last integral is of the form

$$\int_0^\infty\sigma_l^m u^{2m}e^{-\frac{1}{2}u^2}du=\hat{C}_m\sigma_l^m$$

and so, doing this repeatedly, one obtains for the above integral an expression of the form

$$2^p \frac{p^{m-1}}{(2\pi)^{p/2}} C_m \left(\prod_{k=1}^{p} \sigma_k \right)^m = 2^p \frac{p^{m-1}}{(2\pi)^{p/2}} C_m \det(\Sigma)^m$$

which shows that the moments of $\boldsymbol{X}$ all exist and are dominated by an expression which depends continuously on $\det(\Sigma)$. ■

In particular, these moments are bounded in case $\Sigma_n \to \Sigma$ where perhaps $\det(\Sigma) = 0$ but Σ_n is positive definite. With Corollary 28.4.11, this has proved the following theorem about the generalized normal distribution.

Theorem 28.4.14 *Let Σ be nonnegative and self adjoint $p \times p$ matrix. Then there exists a random variable $\boldsymbol{X}$ whose distribution measure $\lambda_{\boldsymbol{X}}$ has characteristic function $\psi(\boldsymbol{t}) \equiv e^{-\frac{1}{2}\boldsymbol{t}^* \Sigma \boldsymbol{t}}$. Then all the moments exist and $E(\boldsymbol{X}\boldsymbol{X}^*) = \Sigma$.*

Proof: It remains to verify $E(\boldsymbol{X}\boldsymbol{X}^*) = \Sigma$ but this is routine from the fact that the moments exist. Use the characteristic function to compute $E(X_i X_j)$. Take $\frac{d}{dt_j}\left(\frac{d}{dt_i}(\psi(\boldsymbol{t}))\right)$. Using repeated index summation convention,

$$\begin{aligned} \psi(\boldsymbol{t}) &= e^{-\frac{1}{2}t_r \Sigma_{rs} t_s}, \psi_{t_i} = e^{-\frac{1}{2}t_r \Sigma_{rs} t_s}(-\Sigma_{is} t_s), \psi_{t_i t_j} \\ &= e^{-\frac{1}{2}t_r \Sigma_{rs} t_s}(-\Sigma_{js} t_s)(-\Sigma_{is} t_s) + e^{-\frac{1}{2}t_r \Sigma_{rs} t_s}(-\Sigma_{ij}) \end{aligned}$$

Thus $i^2 E(X_i X_j) = -\Sigma_{ij}$ showing that $E(\boldsymbol{X}\boldsymbol{X}^*) = \Sigma$ as claimed. ■

The case where $\boldsymbol{m} = \boldsymbol{0}$ is the one of most interest here, but you could always reduce to this case by considering a random variable $\boldsymbol{X} - \boldsymbol{m}$ where $E(\boldsymbol{X}) = \boldsymbol{m}$. There is an interesting corollary to this theorem.

Corollary 28.4.15 *Let H be a real Hilbert space. Then there exist random variables $W(h)$ for $h \in H$ such that for any finite set $\{f_1, f_2, \cdots, f_n\}$,*

$$(W(f_1), W(f_2), \cdots, W(f_n))$$

is normally distributed with mean 0 and covariance $\Sigma_{ij} = (f_i, f_j)$ and for every h, g,

$$E(W(h)W(g)) = (h,g)_H$$

If $\{e_i\}$ is an orthogonal set of vectors of H, then $\{W(e_i)\}$ are independent random variables.

Proof: Let $\mu_{h_1 \cdots h_m}$ be a generalized multivariate normal probability distribution with covariance $\Sigma_{ij} = (h_i, h_j)$ and mean 0. That such a thing exists follows from Theorem 28.4.14. Thus the characteristic function of this probability measure is $e^{-\frac{1}{2}\boldsymbol{t}^* \Sigma \boldsymbol{t}}$. Now consider $E_{k_1} \times \cdots \times E_{k_n}$ for Borel sets E_{k_j} where $\{h_1, \cdots, h_m\} \subseteq \{k_1 \cdots k_n\}$ for $n > m$ and the set $E_{k_j} = \mathbb{R}$ whenever $k_j \notin \{h_1, \cdots, h_m\}$ For simplicity, say $h_1, \cdots, h_m$ are the first m slots of $k_1, \cdots, k_n$. Now consider $\mu_{k_1 \cdots k_n}$,

$$\{h_1 \cdots h_m, k_{m+1} \cdots k_n\} = \{k_1 \cdots k_n\}$$

Let ν be a measure on $\mathscr{B}(\mathbb{R}^m)$ which is given by $\nu(E) \equiv \mu_{k_1 \cdots k_n}(E \times \mathbb{R}^{n-m})$. Then does it follow that $\nu = \mu_{h_1 \cdots h_m}$? If so, then the Kolmogorov consistency condition will hold for

these measures $\mu_{h_1\cdots h_m}$. To determine whether this is so, take the characteristic function of ν. Let Σ_1 be the $n\times n$ matrix which comes from the $\{k_1\cdots k_n\}$ and let Σ_2 be the one which comes from the $\{h_1\cdots h_m\}$.

$$\begin{aligned}\int_{\mathbb{R}^m} e^{i\boldsymbol{t}\cdot\boldsymbol{x}}d\nu(x) &\equiv \int_{\mathbb{R}^{n-m}}\int_{\mathbb{R}^m} e^{i(\boldsymbol{t},\boldsymbol{0})\cdot(\boldsymbol{x},\boldsymbol{y})}d\mu_{k_1\cdots k_n}(x,y)\\ &= e^{-\frac{1}{2}(\boldsymbol{t}^*,\boldsymbol{0}^*)\Sigma_1(\boldsymbol{t},\boldsymbol{0})} = e^{-\frac{1}{2}\boldsymbol{t}^*\Sigma_2\boldsymbol{t}}\end{aligned}$$

which is the characteristic function for $\mu_{h_1\cdots h_m}$. Therefore, these two measures are the same and the Kolmogorov consistency condition holds. It follows from The Kolmogorov extension theorem Theorem 20.3.3 that there exists a measure μ defined on the Borel sets of $\prod_{h\in H}\mathbb{R}$ which extends all of these measures. This argument also shows that if a random vector $\boldsymbol{X}$ has characteristic function $e^{-\frac{1}{2}\boldsymbol{t}^*\Sigma\boldsymbol{t}}$, then if X_k is one of its components, then the characteristic function of X_k is $e^{-\frac{1}{2}t^2|h_k|^2}$ so this scalar valued random variable has mean zero and variance $|h_k|^2$. Then if $\omega\in\prod_{h\in H}\mathbb{R}$, $W(h)(\omega)\equiv\pi_h(\omega)$ where π_h denotes the projection onto position h in this product. Also define

$$(W(f_1),W(f_2),\cdots,W(f_n))\equiv\pi_{f_1\cdots f_n}(\omega)$$

Then this is a random variable whose covariance matrix is just $\Sigma_{ij}=(f_i,f_j)_H$ and whose characteristic equation is $e^{-\frac{1}{2}\boldsymbol{t}^*\Sigma\boldsymbol{t}}$ so this verifies that

$$(W(f_1),W(f_2),\cdots,W(f_n))$$

is normally distributed with covariance Σ. If you have two of them, $W(g),W(h)$, then $E(W(h)W(g))=(h,g)_H$. This follows from what was just shown that $(W(f),W(g))$ is normally distributed and so the covariance will be

$$\begin{pmatrix}|f|^2 & (f,g)\\ (f,g) & |g|^2\end{pmatrix}=\begin{pmatrix}E\left(W(f)^2\right) & E(W(f)W(g))\\ E(W(f)W(g)) & E\left(W(g)^2\right)\end{pmatrix}$$

Finally consider the claim about independence. Any finite subset of $\{W(e_i)\}$ is generalized normal with the covariance matrix being a diagonal. Therefore,

$$(W(e_{i_1}),\cdots,W(e_{i_n}))$$

is normally distributed with covariance a diagonal matrix so by Theorem 28.2.3, the random variables $\{W(e_i)\}$ are independent. ■

28.5 The Central Limit Theorem

The central limit theorem is one of the most marvelous theorems in mathematics. It can be proved through the use of characteristic functions. Recall for $\boldsymbol{x}\in\mathbb{R}^p$,

$$\|\boldsymbol{x}\|_\infty\equiv\max\left\{\left|x_j\right|,j=1,\cdots,p\right\}.$$

Also recall the definition of the distribution function for a random vector, $\boldsymbol{X}$.

$$F_{\boldsymbol{X}}(\boldsymbol{x})\equiv P(X_j\le x_j,j=1,\cdots,p).$$

How can you tell if a sequence of random vectors with values in $\mathbb{R}^p$ is tight? The next lemma gives a way to do this. It is Lemma 28.4.3. I am stating it here for convenience.

Lemma 28.5.1 *If $\{\boldsymbol{X}_n\}$ is a sequence of random vectors with values in $\mathbb{R}^p$ such that*

$$\lim_{n\to\infty}\phi_{\boldsymbol{X}_n}(\boldsymbol{t})\equiv\lim_{n\to\infty}\phi_{\lambda_{\boldsymbol{X}_n}}(\boldsymbol{t})=\psi(\boldsymbol{t})$$

for all $\boldsymbol{t}$, where $\psi(\mathbf{0})=1$ and ψ is continuous at $\mathbf{0}$, then $\{\lambda_{\boldsymbol{X}_n}\}_{n=1}^{\infty}$ is tight.

In proving the central limit theorem, one considers the pointwise convergence of characteristic functions and then seeks to obtain information about the distribution of the limit function. In fact, one is in the situation of the following lemma which is Lemma 28.4.4.

Lemma 28.5.2 *If $\phi_{\boldsymbol{X}_n}(\boldsymbol{t})\to\phi_{\boldsymbol{X}}(\boldsymbol{t})$ for all $\boldsymbol{t}$, then whenever $\psi\in\mathfrak{S}$,*

$$\lambda_{\boldsymbol{X}_n}(\psi)\equiv\int_{\mathbb{R}^p}\psi(\boldsymbol{y})\,d\lambda_{\boldsymbol{X}_n}(y)\to\int_{\mathbb{R}^p}\psi(\boldsymbol{y})\,d\lambda_{\boldsymbol{X}}(y)\equiv\lambda_{\boldsymbol{X}}(\psi)$$

as $n\to\infty$.

The above gives what I want for $\psi\in\mathfrak{S}$ but this needs to be generalized to ψ any bounded uniformly continuous function. The following is Lemma 28.4.5.

Lemma 28.5.3 *If $\phi_{\boldsymbol{X}_n}(\boldsymbol{t})\to\phi_{\boldsymbol{X}}(\boldsymbol{t})$, then if ψ is any bounded uniformly continuous function,*

$$\lim_{n\to\infty}\int_{\mathbb{R}^p}\psi d\lambda_{\boldsymbol{X}_n}=\int_{\mathbb{R}^p}\psi d\lambda_{\boldsymbol{X}}.$$

Definition 28.5.4 *Let μ be a Radon measure on $\mathbb{R}^p$. A Borel set A, is a μ continuity set if $\mu(\partial A)=0$ where $\partial A\equiv\overline{A}\setminus\operatorname{int}(A)$ and* int *denotes the interior.*

The main result is the following continuity theorem. More can be said about the equivalence of various criteria [6].

Theorem 28.5.5 *If $\phi_{\boldsymbol{X}_n}(\boldsymbol{t})\to\phi_{\boldsymbol{X}}(\boldsymbol{t})$ then $\lambda_{\boldsymbol{X}_n}(A)\to\lambda_{\boldsymbol{X}}(A)$ whenever A is a $\lambda_{\boldsymbol{X}}$ continuity set.*

Proof: First suppose K is a closed set and let

$$\psi_k(\boldsymbol{x})\equiv(1-k\operatorname{dist}(\boldsymbol{x},K))^+.$$

Thus, since K is closed $\lim_{k\to\infty}\psi_k(\boldsymbol{x})=\mathcal{X}_K(\boldsymbol{x})$. Choose k large enough that

$$\int_{\mathbb{R}^p}\psi_k d\lambda_{\boldsymbol{X}}\le\lambda_{\boldsymbol{X}}(K)+\varepsilon.$$

Then by Lemma 28.5.3, applied to the bounded uniformly continuous function ψ_k,

$$\limsup_{n\to\infty}\lambda_{\boldsymbol{X}_n}(K)\le\limsup_{n\to\infty}\int\psi_k d\lambda_{\boldsymbol{X}_n}=\int\psi_k d\lambda_{\boldsymbol{X}}\le\lambda_{\boldsymbol{X}}(K)+\varepsilon.$$

Since ε is arbitrary, this shows $\limsup_{n\to\infty}\lambda_{\boldsymbol{X}_n}(K)\le\lambda_{\boldsymbol{X}}(K)$ for all K closed.

Next suppose V is open and let

$$\psi_k(\boldsymbol{x})=1-\left(1-k\operatorname{dist}\left(\boldsymbol{x},V^C\right)\right)^+.$$

Thus $\psi_k(\boldsymbol{x}) \in [0,1], \psi_k = 1$ if $\text{dist}(\boldsymbol{x}, V^C) \geq 1/k$, and $\psi_k = 0$ on V^C. Since V is open, it follows $\lim_{k\to\infty} \psi_k(\boldsymbol{x}) = \mathcal{X}_V(\boldsymbol{x})$. Choose k large enough that $\int \psi_k d\lambda_{\boldsymbol{X}} \geq \lambda_{\boldsymbol{X}}(V) - \varepsilon$. Then by Lemma 28.5.3,

$$\liminf_{n\to\infty} \lambda_{\boldsymbol{X}_n}(V) \geq \liminf_{n\to\infty} \int \psi_k(\boldsymbol{x})\, d\lambda_{\boldsymbol{X}_n} = \int \psi_k(\boldsymbol{x})\, d\lambda_{\boldsymbol{X}} \geq \lambda_{\boldsymbol{X}}(V) - \varepsilon$$

and since ε is arbitrary, $\liminf_{n\to\infty} \lambda_{\boldsymbol{X}_n}(V) \geq \lambda_{\boldsymbol{X}}(V)$. Now let $\lambda_{\boldsymbol{X}}(\partial A) = 0$ for A a Borel set.

$$\begin{aligned} \lambda_{\boldsymbol{X}}(\text{int}(A)) &\leq \liminf_{n\to\infty} \lambda_{\boldsymbol{X}_n}(\text{int}(A)) \leq \liminf_{n\to\infty} \lambda_{\boldsymbol{X}_n}(A) \leq \\ \limsup_{n\to\infty} \lambda_{\boldsymbol{X}_n}(A) &\leq \limsup_{n\to\infty} \lambda_{\boldsymbol{X}_n}(\overline{A}) \leq \lambda_{\boldsymbol{X}}(\overline{A}). \end{aligned}$$

But $\lambda_{\boldsymbol{X}}(\text{int}(A)) = \lambda_{\boldsymbol{X}}(\overline{A})$ by assumption and so $\lim_{n\to\infty} \lambda_{\boldsymbol{X}_n}(A) = \lambda_{\boldsymbol{X}}(A)$ as claimed. ■

As an application of this theorem the following is a version of the central limit theorem in the situation in which the limit distribution is multivariate normal. It concerns a sequence of random vectors, $\{\boldsymbol{X}_k\}_{k=1}^{\infty}$, which are identically distributed, have finite mean $\boldsymbol{m}$, and satisfy $E\left(|\boldsymbol{X}_k|^2\right) < \infty$.

Definition 28.5.6 *For $\boldsymbol{X}$ a random vector with values in $\mathbb{R}^p$, let*

$$F_{\boldsymbol{X}}(\boldsymbol{x}) \equiv P\left(\{X_j \leq x_j \text{ for each } j = 1,2,...,p\}\right).$$

A different proof of the central limit theorem is in [47].

Lemma 28.5.7 *If all the z_i and w_i have absolute value no more than 1, then $\left|\prod_{i=1}^n z_i - \prod_{i=1}^n w_i\right| \leq \sum_{k=1}^n |z_k - w_k|$.*

Proof: It is clearly true if $n = 1$. Suppose true for n. Then

$$\left|\prod_{i=1}^{n+1} z_i - \prod_{i=1}^{n+1} w_i\right| \leq \left|\prod_{i=1}^{n+1} z_i - z_{n+1}\prod_{i=1}^{n} w_i\right| + \left|z_{n+1}\prod_{i=1}^{n} w_i - \prod_{i=1}^{n+1} w_i\right|$$

$$\leq \left|\prod_{i=1}^{n} z_i - \prod_{i=1}^{n} w_i\right| + |z_{n+1} - w_{n+1}|\left|\prod_{i=1}^{n} w_i\right| \leq \sum_{k=1}^{n+1} |z_k - w_k| \blacksquare$$

Theorem 28.5.8 *Let $\{\boldsymbol{X}_k\}_{k=1}^{\infty}$ be random vectors satisfying $E\left(|\boldsymbol{X}_k|^2\right) < \infty$ which are independent and identically distributed with mean $\boldsymbol{m}$ and positive definite covariance $\boldsymbol{\Sigma} \equiv E\left((\boldsymbol{X} - \boldsymbol{m})(\boldsymbol{X} - \boldsymbol{m})^*\right)$. Let $\boldsymbol{Z}_n \equiv \sum_{j=1}^n \frac{\boldsymbol{X}_j - \boldsymbol{m}}{\sqrt{n}}$. Then for $\boldsymbol{Z} \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$, $\lim_{n\to\infty} F_{\boldsymbol{Z}_n}(\boldsymbol{x}) = F_{\boldsymbol{Z}}(\boldsymbol{x})$ for all $\boldsymbol{x}$.*

Proof: The characteristic function of $\boldsymbol{Z}_n$ is given by

$$\phi_{\boldsymbol{Z}_n}(\boldsymbol{t}) = E\left(e^{i\boldsymbol{t}\cdot\sum_{j=1}^n \frac{\boldsymbol{X}_j - \boldsymbol{m}}{\sqrt{n}}}\right) = \prod_{j=1}^{n} E\left(e^{i\boldsymbol{t}\cdot\left(\frac{\boldsymbol{X}_j - \boldsymbol{m}}{\sqrt{n}}\right)}\right).$$

By Taylor's theorem applied to real and imaginary parts of e^{ix}, it follows

$$e^{ix} = 1 + ix - f(x)\frac{x^2}{2}$$

where $|f(x)| < 2$ and $\lim_{x\to 0} f(x) = 1$. Denoting $\boldsymbol{X}_j$ as $\boldsymbol{X}$, this implies

$$e^{i\boldsymbol{t}\cdot\left(\frac{\boldsymbol{X}-\boldsymbol{m}}{\sqrt{n}}\right)} = 1 + i\,\boldsymbol{t}\cdot\frac{\boldsymbol{X}-\boldsymbol{m}}{\sqrt{n}} - f\left(\boldsymbol{t}\cdot\left(\frac{\boldsymbol{X}-\boldsymbol{m}}{\sqrt{n}}\right)\right)\frac{(\boldsymbol{t}\cdot(\boldsymbol{X}-\boldsymbol{m}))^2}{2n}$$

Thus $e^{i\boldsymbol{t}\cdot\left(\frac{\boldsymbol{X}-\boldsymbol{m}}{\sqrt{n}}\right)} = 1 + i\,\boldsymbol{t}\cdot\frac{\boldsymbol{X}-\boldsymbol{m}}{\sqrt{n}} - \frac{(\boldsymbol{t}\cdot(\boldsymbol{X}-\boldsymbol{m}))^2}{2n} + \left(1 - f\left(\boldsymbol{t}\cdot\left(\frac{\boldsymbol{X}-\boldsymbol{m}}{\sqrt{n}}\right)\right)\right)\frac{(\boldsymbol{t}\cdot(\boldsymbol{X}-\boldsymbol{m}))^2}{2n}$. This implies

$$\phi_{\boldsymbol{Z}_n}(\boldsymbol{t}) = \prod_{j=1}^{n} E\left[1 - \frac{(\boldsymbol{t}\cdot(\boldsymbol{X}_j-\boldsymbol{m}))^2}{2n} + \frac{(\boldsymbol{t}\cdot(\boldsymbol{X}_j-\boldsymbol{m}))^2}{2n}\left(1 - f\left(\boldsymbol{t}\cdot\left(\frac{\boldsymbol{X}_j-\boldsymbol{m}}{\sqrt{n}}\right)\right)\right)\right]$$

Then $\phi_{\boldsymbol{Z}_n}(\boldsymbol{t}) =$

$$\prod_{j=1}^{n} E\left[1 - \frac{(\boldsymbol{t}\cdot(\boldsymbol{X}_j-\boldsymbol{m}))^2}{2n} + \frac{(\boldsymbol{t}\cdot(\boldsymbol{X}_j-\boldsymbol{m}))^2}{2n}\left(1 - f\left(\boldsymbol{t}\cdot\left(\frac{\boldsymbol{X}_j-\boldsymbol{m}}{\sqrt{n}}\right)\right)\right)\right]$$
$$-\prod_{j=1}^{n} E\left[1 - \frac{(\boldsymbol{t}\cdot(\boldsymbol{X}_j-\boldsymbol{m}))^2}{2n}\right] + \prod_{j=1}^{n}\left(1 - \frac{E(\boldsymbol{t}\cdot(\boldsymbol{X}_j-\boldsymbol{m}))^2}{2n}\right)$$

Now $(\boldsymbol{t}\cdot(\boldsymbol{X}-\boldsymbol{m}))^2 = \boldsymbol{t}^*(\boldsymbol{X}-\boldsymbol{m})(\boldsymbol{X}-\boldsymbol{m})^*\boldsymbol{t}$. Since these $\boldsymbol{X}_k$ are identically distributed with the same mean $\boldsymbol{m}$, the above is of the form

$$e_n + \prod_{j=1}^{n}\left(1 - \frac{E(\boldsymbol{t}\cdot(\boldsymbol{X}_j-\boldsymbol{m}))^2}{2n}\right) = e_n + \left(1 - \frac{1}{2n}\boldsymbol{t}^*\Sigma\boldsymbol{t}\right)^n$$

where for large n, the needed expressions have small absolute value and so, from the above lemma, for large n,

$$|e_n| \le \frac{1}{2n}\sum_{j=1}^{n} E\left((\boldsymbol{t}\cdot(\boldsymbol{X}_j-\boldsymbol{m}))^2\left|1 - f\left(\boldsymbol{t}\cdot\left(\frac{\boldsymbol{X}_j-\boldsymbol{m}}{\sqrt{n}}\right)\right)\right|\right)$$

Now write $\boldsymbol{X}$ for $\boldsymbol{X}_k$ since all are identically distributed. Then the above right side is no more than

$$\frac{1}{2}E\left((\boldsymbol{t}\cdot(\boldsymbol{X}-\boldsymbol{m}))^2\left|1 - f\left(\boldsymbol{t}\cdot\left(\frac{\boldsymbol{X}-\boldsymbol{m}}{\sqrt{n}}\right)\right)\right|\right)$$

which converges to 0 as $n\to\infty$ by the dominated convergence theorem. Therefore,

$$\lim_{n\to\infty}\phi_{\boldsymbol{Z}_n}(\boldsymbol{t}) = \lim_{n\to\infty}\left(1 - \frac{1}{2n}\boldsymbol{t}^*\Sigma\boldsymbol{t}\right)^n = e^{-\frac{1}{2}\boldsymbol{t}^*\Sigma\boldsymbol{t}} = \phi_{\boldsymbol{Z}}(\boldsymbol{t})$$

where $\boldsymbol{Z} \sim N_p(\boldsymbol{0}, \Sigma)$. Therefore, from Theorem 28.5.5, $F_{\boldsymbol{Z}_n}(\boldsymbol{x}) \to F_{\boldsymbol{Z}}(\boldsymbol{x})$ for all $\boldsymbol{x}$ because

$$R_{\boldsymbol{x}} \equiv \prod_{k=1}^{p}(-\infty, x_k]$$

is a set of $\lambda_{\boldsymbol{Z}}$ continuity due to the assumption that $\lambda_{\boldsymbol{Z}} \ll m_p$ which is implied by $\boldsymbol{Z} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$. ■

Suppose $\boldsymbol{X}$ is a random vector with covariance Σ and mean $\boldsymbol{m}$, and suppose also that Σ^{-1} exists. Consider $\Sigma^{-(1/2)}(\boldsymbol{X}-\boldsymbol{m}) \equiv \boldsymbol{Y}$. Then $E(\boldsymbol{Y}) = 0$ and

$$\begin{aligned} E(\boldsymbol{Y}\boldsymbol{Y}^*) &= E\left(\Sigma^{-(1/2)}(\boldsymbol{X}-\boldsymbol{m})(\boldsymbol{X}^*-\boldsymbol{m})\Sigma^{-(1/2)}\right) \\ &= \Sigma^{-(1/2)}E((\boldsymbol{X}-\boldsymbol{m})(\boldsymbol{X}^*-\boldsymbol{m}))\Sigma^{-(1/2)} = I. \end{aligned}$$

Thus $\boldsymbol{Y}$ has zero mean and covariance I. This implies the following corollary to Theorem 28.5.8.

Corollary 28.5.9 *Let $\{\boldsymbol{X}_j\}_{j=1}^{\infty}$ be independent identically distributed random variables and suppose they have mean $\boldsymbol{m}$ and positive definite covariance $\boldsymbol{\Sigma}$ where $\boldsymbol{\Sigma}^{-1}$ exists. Then if*

$$\boldsymbol{Z}_n \equiv \sum_{j=1}^{n} \boldsymbol{\Sigma}^{-(1/2)} \frac{(\boldsymbol{X}_j - \boldsymbol{m})}{\sqrt{n}},$$

it follows that for $\boldsymbol{Z} \sim N_p(\mathbf{0}, I)$, $F_{\boldsymbol{Z}_n}(\boldsymbol{x}) \to F_{\boldsymbol{Z}}(\boldsymbol{x})$ for all $\boldsymbol{x}$.

Chapter 29

Martingales

29.1 Conditional Expectation

From Observation 27.4.5 on Page 777, it was shown that the conditional expectation of a random variable $\boldsymbol{X}$ given some others really is just what the words suggest. Given $\omega \in \Omega$, it results in a value for the "other" random variables and then you essentially take the expectation of $\boldsymbol{X}$ given this information which yields the value of the conditional expectation of $\boldsymbol{X}$ given the other random variables. It was also shown in Lemma 27.4.4 that this gives the same result as finding a $\sigma(\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)$ measurable function $\boldsymbol{Z}$ such that for all $F \in \sigma(\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)$,

$$\int_F \boldsymbol{X} dP = \int_F \boldsymbol{Z} dP$$

This was done for a particular type of σ algebra but there is no need to be this specialized. The following is the general version of conditional expectation given a σ algebra. It makes perfect sense to ask for the conditional expectation given a σ algebra and this is what will be done from now on.

Definition 29.1.1 *Let $(\Omega, \mathscr{M}, P)$ be a probability space and let $\mathscr{S} \subseteq \mathscr{F}$ be two σ algebras contained in $\mathscr{M}$. Let f be $\mathscr{F}$ measurable and in $L^1(\Omega; W)$ for W a Banach space. Then $E(f|\mathscr{S})$, called the conditional expectation of f with respect to $\mathscr{S}$ is defined as follows:*

$$E(f|\mathscr{S}) \text{ is } \mathscr{S} \text{ measurable}$$

For all $E \in \mathscr{S}$,

$$\int_E E(f|\mathscr{S}) dP = \int_E f dP$$

The existence and uniqueness of the conditional expectation is described earlier in Theorem 24.12.1 on Page 729. For convenience, here is this theorem.

Theorem 29.1.2 *Let E be a separable Banach space and suppose $X \in L^1(\Omega; E, \mathscr{M})$ where X is measurable with respect to $\mathscr{M}$ and let $\mathscr{S}$ be a σ algebra which is contained in $\mathscr{M}$. Then there exists a unique $Z \in L^1(\Omega; E, \mathscr{S})$ such that for all $A \in \mathscr{S}$,*

$$\int_A X dP = \int_A Z dP$$

Denoting this Z as $E(X|\mathscr{S})$, it follows

$$\|E(X|\mathscr{S})\| \leq E(\|X\| \ |\mathscr{S}).$$

A few properties are described next. Let W be a separable Banach space in the following lemma.

Lemma 29.1.3 *The above is well defined. Also, if $\mathscr{S} \subseteq \mathscr{F}$ then if $X \in L^1(\Omega; W)$,*

$$E(X|\mathscr{S}) = E(E(X|\mathscr{F})|\mathscr{S}). \tag{29.1}$$

If Z is in $L^\infty(\Omega; W')$ bounded and measurable in $\mathscr{S}$ then

$$ZE(X|\mathscr{S}) = E(ZX|\mathscr{S}). \tag{29.2}$$

Also, if $a,b \in W'$, *and* $X,Y \in L^1(\Omega;W)$

$$aE(X|\mathcal{S}) + bE(Y|\mathcal{S}) = E(aX+bY|\mathcal{S}). \tag{29.3}$$

Proof: To begin with consider 29.3. By definition, if $F \in \mathcal{S}$,

$$\int_F aE(X|\mathcal{S}) + bE(Y|\mathcal{S})\,dP = a\int_F E(X|\mathcal{S})\,dP + b\int_F E(Y|\mathcal{S})\,dP$$

$$= a\int_F X\,dP + b\int_F Y\,dP = \int_F (aX+bY)\,dP \equiv \int_F E(aX+bY|\mathcal{S})\,dP$$

Since F is arbitrary, this shows 29.3.

Let $F \in \mathcal{S}$. Then

$$\begin{aligned}\int_F E(E(X|\mathcal{F})|\mathcal{S})\,dP &\equiv \int_F E(X|\mathcal{F})\,dP \\ &\equiv \int_F X\,dP \equiv \int_F E(X|\mathcal{S})\,dP\end{aligned}$$

and so, by uniqueness, $E(E(X|\mathcal{F})|\mathcal{S}) = E(X|\mathcal{S})$. This shows 29.1.

To establish 29.2, note that if $Z = a\mathcal{X}_F$ where $F \in \mathcal{S}$, and $a \in W'$, by Definition 29.1.1,

$$\begin{aligned}\int a\mathcal{X}_F E(X|\mathcal{S})\,dP &= \int_F E(aX|\mathcal{S})\,dP\,dP = \int_F aX\,dP \\ &= \int a\mathcal{X}_F X\,dP = \int E(a\mathcal{X}_F X|\mathcal{S})\,dP\end{aligned}$$

which shows 29.2 in the case where Z is $a\mathcal{X}_F, F \in \mathcal{S}$. It follows this also holds for simple functions with values in W'. Let Z be in $L^\infty(\Omega;W)$. By Theorem 24.2.4 there is a sequence of simple functions $\{s_n\}$, $\|s_n(\omega)\| \le 2\|Z(\omega)\|$ which converges to Z and let $F \in \mathcal{S}$. Then by what was just shown,

$$\int_F s_n E(X|\mathcal{S})\,dP = \int_F E(s_n X|\mathcal{S})\,dP \equiv \int_F s_n X\,dP \tag{29.4}$$

Now

$$\begin{aligned}\left\|\int_F E(s_n X|\mathcal{S})\,dP - \int_F E(ZX|\mathcal{S})\,dP\right\| &= \left\|\int_F (s_n - Z)X\,dP\right\| \\ &\le \int_F \|(s_n - Z)X\|\,dP\end{aligned}$$

and this converges to 0 by the dominated convergence theorem. Also from Theorem 24.12.1

$$\|s_n E(X|\mathcal{S})\| = \|E(s_n X|\mathcal{S})\| \le E(\|s_n X\||\mathcal{S}) \le 2E(\|ZX\||\mathcal{S})$$

which is in $L^1(\Omega)$. Thus one can apply the dominated convergence theorem to the left side of 29.4 and use what was just shown to pass to a limit in 29.4 and obtain

$$\int_F ZE(X|\mathcal{S})\,dP = \int_F ZX\,dP \equiv \int_F E(ZX|\mathcal{S})\,dP.$$

Since this holds for every $F \in \mathcal{S}$, this shows 29.2. ■

The next major result is a generalization of Jensen's inequality whose proof depends on the following lemma about convex functions. It pertains to the case where the functions have values in $\mathbb{R}$.

Lemma 29.1.4 *Let ϕ be a convex real valued function defined on an interval I. Then for each $x \in I$, there exists a_x such that for all $t \in I$,*

$$\phi(t) \geq a_x(t-x) + \phi(x).$$

Also ϕ is continuous on I.

Proof: Let $x \in I$ and let $t > x$. Then by convexity of ϕ,

$$\frac{\phi(x+\lambda(t-x)) - \phi(x)}{\lambda(t-x)} \leq \frac{\phi(x)(1-\lambda) + \lambda\phi(t) - \phi(x)}{\lambda(t-x)} = \frac{\phi(t) - \phi(x)}{t-x}.$$

Therefore $t \to \frac{\phi(t)-\phi(x)}{t-x}$ is increasing if $t > x$. If $t < x, t-x < 0$ so

$$\frac{\phi(x+\lambda(t-x)) - \phi(x)}{\lambda(t-x)} \geq \frac{\phi(x)(1-\lambda) + \lambda\phi(t) - \phi(x)}{\lambda(t-x)} = \frac{\phi(t) - \phi(x)}{t-x}$$

and so $t \to \frac{\phi(t)-\phi(x)}{t-x}$ is increasing for $t \neq x$. Let

$$a_x \equiv \inf\left\{\frac{\phi(t) - \phi(x)}{t-x} : t > x\right\}.$$

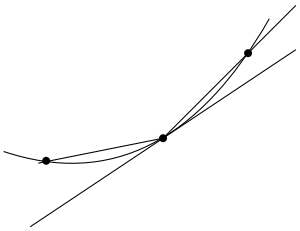
Then if $t_1 < x$, and $t > x$,

$$\frac{\phi(t_1) - \phi(x)}{t_1 - x} \leq a_x \leq \frac{\phi(t) - \phi(x)}{t-x}.$$

Thus for all $t \in I$,

$$\phi(t) \geq a_x(t-x) + \phi(x). \tag{29.5}$$

The continuity of ϕ follows easily from this and the observation that convexity simply says that the graph of ϕ lies below the line segment joining two points on its graph. Thus, we have the following picture which clearly implies continuity. ■



Lemma 29.1.5 *Let I be an interval on $\mathbb{R}$ and let ϕ be a convex function defined on I. Then there exists a sequence $\{(a_n, b_n)\}$ such that*

$$\phi(t) = \sup\{a_n t + b_n, n = 1, \cdots\}.$$

Proof: Let a_x be as defined in the above lemma. Let

$$\psi(x) \equiv \sup\{a_r(x-r) + \phi(r) : r \in \mathbb{Q} \cap I\}.$$

Thus if $r_1 \in \mathbb{Q}$, $\psi(r_1) \equiv \sup\{a_r(r_1 - r) + \phi(r) : r \in \mathbb{Q} \cap I\} \geq \phi(r_1)$. Then ψ is convex on I so ψ is continuous. Therefore, $\psi(t) \geq \phi(t)$ for all $t \in I$. By 29.5,

$$\psi(t) \geq \phi(t) \geq \sup\{a_r(t - r) + \phi(r), r \in \mathbb{Q} \cap I\} \equiv \psi(t).$$

Thus $\psi(t) = \phi(t)$. Let $\mathbb{Q} \cap I = \{r_n\}$, $a_n = a_{r_n}$ and $b_n = -a_{r_n} r_n + \phi(r_n)$. Continuity gives the desired results at endpoints of I. ■

In this lemma, X, Y have values in $\mathbb{R}$.

Lemma 29.1.6 *If $X \leq Y$, then $E(X|\mathcal{S}) \leq E(Y|\mathcal{S})$ a.e. Also $X \to E(X|\mathcal{S})$ is linear.*

Proof: Let $A \in \mathcal{S}$.

$$\int_A E(X|\mathcal{S})\, dP \equiv \int_A X dP \leq \int_A Y dP \equiv \int_A E(Y|\mathcal{S})\, dP.$$

Hence $E(X|\mathcal{S}) \leq E(Y|\mathcal{S})$ a.e. as claimed. That $X \to E(X|\mathcal{S})$ is linear follows from Lemma 29.1.3.

Theorem 29.1.7 *(Jensen's inequality)Let $X(\omega) \in I$ a closed interval and let $\phi : I \to \mathbb{R}$ be convex. Suppose $E(|X|), E(|\phi(X)|) < \infty$. Then $\phi(E(X|\mathcal{S})) \leq E(\phi(X)|\mathcal{S})$.*

Proof: Let $\phi(x) = \sup\{a_n x + b_n\}$. Letting $A \in \mathcal{S}$,

$$\frac{1}{P(A)} \int_A E(X|\mathcal{S})\, dP = \frac{1}{P(A)} \int_A X dP \in I \text{ a.e.}$$

whenever $P(A) \neq 0$. The claim that $\frac{1}{P(A)} \int_A X dP \in I$ follows from approximating X with simple functions having values in I. Hence $E(X|\mathcal{S})(\omega) \in I$ a.e. and so it makes sense to consider $\phi(E(X|\mathcal{S}))$. Now $a_n E(X|\mathcal{S}) + b_n = E(a_n X + b_n|\mathcal{S}) \leq E(\phi(X)|\mathcal{S})$.Thus $\sup\{a_n E(X|\mathcal{S}) + b_n\} = \phi(E(X|\mathcal{S})) \leq E(\phi(X)|\mathcal{S})$ a.e. ■

29.2 Conditional Expectation and Independence

The situation of interest is a sequence of random variables $\{Y_i\}$ having values in a separable Banah space along with two other random variables X, Z both of which are measurable with respect to $\mathcal{E}$ where $\mathcal{E}, \sigma(Y_1, ...)$ are independent σ algebras contained in $\mathcal{M}$, so if $A \in \mathcal{E}$ and $B \in \sigma(Y_1, ...)$, then $P(A \cap B) = P(A)P(B)$. Also let $\sigma(Z, \mathbf{Y})$ be the smallest σ algebra for which Z and each Y_k are measurable. Then both X, Z would seem to relate only to $\mathcal{E}$ and so it would seem that the values of the Y_k would be irrelevant and $E(X|Z) = E(X|\sigma(Z, \mathbf{Y}))$. That is, the conditional expectation given the extra conditions from the Y_k is unchanged. Is it like the earlier notion in which independence means you can dispense with the givens? Recall also the notation $E(X|Z)$ is defined as $E(X|\sigma(Z))$.

Recall that if $f^{-1}(O) \in \mathcal{F}$ a σ algebra and this holds for all O open, then if $\mathcal{S} \equiv \left\{B : f^{-1}(B) \in \mathcal{F}\right\}$ it follows that $\mathcal{S}$ is a σ algebra and so it contains the Borel sets.

Proposition 29.2.1 *Let $\mathcal{E}, \sigma(Y_1, Y_2...)$ be independent σ algebras contained in $\mathcal{M}$. Also let $\sigma(Z, \mathbf{Y})$ be the smallest σ algebra which respect to which each Y_k and Z is measurable. Let Z, X be $\mathcal{E}$ measurable. Then $E(X|Z) = E(X|\sigma(Z, \mathbf{Y}))$.*

Proof: First I claim that $\sigma(Z,\boldsymbol{Y})$ consists of $(Z,\boldsymbol{Y})^{-1}(B)$ where B is Borel in $W\times W^{\mathbb{N}}$ where $W^{\mathbb{N}}\equiv\prod_{i=1}^{\infty}W$. To see this, let

$$\left\{(Z,\boldsymbol{Y})^{-1}(B) : B \text{ is Borel in } W^{1+\mathbb{N}}\right\}\equiv\mathscr{F}.$$

Then $\mathscr{F}$ is a σ algebra and for V_J of the form $\prod_{j=1}^{\infty}V_j$ where $V_j=W$ for all j except for those contained in a finite set J and for the other j, V_j is an open set, and U an open set in W, then $(Z,\boldsymbol{Y})^{-1}(U\times V_J)=Z^{-1}(U)\cap\boldsymbol{Y}^{-1}(V_J)\in\mathscr{F}$ so that, in particular, choosing V_J and U appropriately shows Z, Y_k are all measurable with respect to $\mathscr{F}$. Hence $\mathscr{F}\supseteq\sigma(Z,\boldsymbol{Y})$. Also, $U\times V_J$ just described, where U is in a countable basis for W and each V_j is W or in a countable basis for W is a countable basis for the topology of $W^{1+\mathbb{N}}$. By definition, $\sigma(Z,\boldsymbol{Y})$ must contain $(Z,\boldsymbol{Y})^{-1}(U\times V_J)$ for $U\times V_J$ in this countable basis. Therefore, $\sigma(Z,\boldsymbol{Y})$ must contain $(Z,\boldsymbol{Y})^{-1}(O)$ for all O open and so also $\sigma(Z,\boldsymbol{Y})$ must contain $(Z,\boldsymbol{Y})^{-1}(B)$ for B Borel, so $\sigma(Z,\boldsymbol{X})\supseteq\mathscr{F}$. Also, this shows that for $\mathscr{K}$ these sets in the countable basis for $W^{1+\mathbb{N}}$, $\sigma(\mathscr{K})=\sigma(Z,\boldsymbol{Y})=\mathscr{F}$. Now consider the following computation.

$$\int_{(Z,\boldsymbol{Y})^{-1}(U\times V_J)}E(X|\sigma(Z,\boldsymbol{Y}))\,dP$$

$$\begin{aligned}
&\equiv \int_{(Z,\boldsymbol{Y})^{-1}(U\times V_J)}X\,dP\\
&= \int\mathscr{X}_{\boldsymbol{Y}^{-1}(V_J)}\mathscr{X}_{Z^{-1}(U)}X\,dP\overset{*}{=}P\left(\boldsymbol{Y}^{-1}(V_J)\right)\int\mathscr{X}_{Z^{-1}(U)}X\,dP\\
&\overset{**}{=}P\left(\boldsymbol{Y}^{-1}(V_J)\right)\int E\left(\mathscr{X}_{Z^{-1}(U)}X|Z\right)dP=\int\mathscr{X}_{\boldsymbol{Y}^{-1}(V_J)}E\left(\mathscr{X}_{Z^{-1}(U)}X|Z\right)dP\\
&=\int_{Z^{-1}(U)}\mathscr{X}_{\boldsymbol{Y}^{-1}(V_J)}E(X|Z)\,dP=\int_{(Z,\boldsymbol{Y})^{-1}(U\times V_J)}E(X|Z)\,dP
\end{aligned}$$

Then $*$ happens because $\mathscr{X}_{[Z\in U]}X$ is $\mathscr{E}$ measurable and $\mathscr{X}_{\boldsymbol{Y}^{-1}(V_J)}$ is $\mathscr{F}$ measurable and by assumption, these are independent σ algebras. $**$ happens because Ω is $\sigma(Z)$ measurable and the definition of conditional expectation. Next step happens because $\mathscr{X}_{\boldsymbol{Y}^{-1}(V_J)}$ is $\mathscr{F}$ measurable and $E\left(\mathscr{X}_{Z^{-1}(U)}X|Z\right)$ is $\mathscr{E}$ measurable. Then the rest follows because $\mathscr{X}_{Z^{-1}(U)}$ is $\sigma(Z)$ measurable so it comes out of the conditional expectation.

Now let $\mathscr{G}$ be those sets G of $\sigma(Z,\boldsymbol{Y})=\sigma(\mathscr{K})$ such that

$$\int_G X\,dP\equiv\int_G E(X|\sigma(Z,\boldsymbol{Y}))\,dP=\int_G E(X|\sigma(Z))\,dP\equiv\int_G E(X|Z)\,dP$$

As just shown, $\mathscr{K}\subseteq\mathscr{G}$. If $G=\cup_{k=1}^{\infty}G_k$ where the G_k are disjoint and the equation holds for each G_k

$$\begin{aligned}
\int_G E(X|Z)\,dP &= \int\sum_{k=1}^{\infty}\mathscr{X}_{G_k}E(X|Z)\,dP=\sum_{k=1}^{\infty}\int_{G_k}E(X|Z)\,dP\\
&= \sum_{k=1}^{\infty}\int_{G_k}E(X|\sigma(Z,\boldsymbol{Y}))\,dP=\int\sum_{k=1}^{\infty}\mathscr{X}_{G_k}E(X|\sigma(Z,\boldsymbol{Y}))\,dP\\
&= \int_G E(X|\sigma(Z,\boldsymbol{Y}))\,dP
\end{aligned}$$

Thus $\mathscr{G}$ is closed with respect to countable disjoint unions. If $G \in \mathscr{G}$ then both sides of the following equal $E(X)$ and so

$$\begin{aligned} & \int_G E(X|\sigma(Z,\boldsymbol{Y}))\,dP + \int_{G^C} E(X|\sigma(Z,\boldsymbol{Y}))\,dP \\ = & \int_G E(X|Z)\,dP + \int_{G^C} E(X|Z)\,dP \end{aligned}$$

and subtracting $\int_G E(X|\sigma(Z,\boldsymbol{Y}))\,dP = \int_G E(X|Z)\,dP$ from both sides yields

$$\int_{G^C} E(X|\sigma(Z,\boldsymbol{Y}))\,dP = \int_{G^C} E(X|Z)\,dP.$$

This proves $\mathscr{G}$ is all of $\sigma(\mathscr{K}) = \sigma(Z,\boldsymbol{Y})$. By uniqueness, $E(X|Z) = E(X|\sigma(Z,\boldsymbol{Y}))$ a.e. ■

29.3 Discrete Stochastic Processes

Earlier a special case of a discreet martingale and sub-martingale was discussed. This section considers the general case where one just has an increasing list of σ algebras. The idea is that you have an increasing list of real numbers $\{a_n\}$ which is well ordered and $X(a_n) \equiv X_{a_n}$ is measurable with respect to $\mathscr{F}_{a_n}$ where the $\mathscr{F}_{a_n}$ are increasing in n. Such a sequence of $\mathscr{F}_{a_n}$ measurable functions is called a stochastic processes.We usually let this well ordered increasing list be some subset of $\mathbb{N}$ for the sake of convenience, but it could be equally spaced points in an interval, for example. We let $\mathscr{F}$ denote a σ algebra which contains all of these $\mathscr{F}_k$. For convenience, you could call it $\mathscr{F}_\infty$.

Definition 29.3.1 *Let $\mathscr{F}_k$ be an increasing sequence of σ algebras which are subsets of $\mathscr{F}$ and X_k be a sequence of Banach space valued random variables with $E(\|X_k\|) < \infty$ such that X_k is $\mathscr{F}_k$ measurable. Such a thing is called a stochastic process. It is called a martingale if*

$$E(X_{k+1}|\mathscr{F}_k) = X_k,$$

In case the Banach space is $\mathbb{R}$ the stochastic process is a sub-martingale if

$$E(X_{k+1}|\mathscr{F}_k) \geq X_k,$$

and a supermartingale if

$$E(X_{k+1}|\mathscr{F}_k) \leq X_k.$$

Saying that X_k is $\mathscr{F}_k$ measurable is referred to by saying $\{X_k\}$ is adapted to $\mathscr{F}_k$.

For sub and super martingales, you need to be considering X which has values in $\mathbb{R}$. No such restriction is necessary for a martingale. If $\{X_k\}$ is a Banach space valued martingale, then from Theorem 29.1.2, $\{\|X_k\|\}$ is a sub-martingale and that if $\{X_k\}$ is a real sub-martingale and ϕ is convex and **increasing,** then $\{\phi(X_k)\}$ is a sub-martingale. This is discussed below.

Also in general, for a stochastic process, $E(X_n|\mathscr{F}_{n-2}) = E(E(X_n|\mathscr{F}_{n-1})|\mathscr{F}_{n-2})$. Thus if $\{X_n\}$ is a sub-martingale, Lemma 29.1.6 implies $E(X_n|\mathscr{F}_{n-2}) \geq X_{n-2}$. Similarly,

$$S(X_n|\mathscr{F}_k) \geq X_k$$

whenever $k < n$. Something similar happens with a martingale where you can replace $\geq$ with $=$ or a super- martingale where you replace $\geq$ with $\leq$. I will use this observation without comment in what follows.

Lemma 29.3.2 *Let ϕ be a convex and increasing function and suppose*

$$\{(X_n, \mathscr{F}_n)\}$$

is a sub-martingale. Then if $E(|\phi(X_n)|) < \infty$, it follows

$$\{(\phi(X_n), \mathscr{F}_n)\}$$

is also a sub-martingale.

Proof: It is given that $E(X_{n+1}, \mathscr{F}_n) \geq X_n$ and so

$$\phi(X_n) \leq \phi(E(X_{n+1}|\mathscr{F}_n)) \leq E(\phi(X_{n+1})|\mathscr{F}_n)$$

by Jensen's inequality. ■

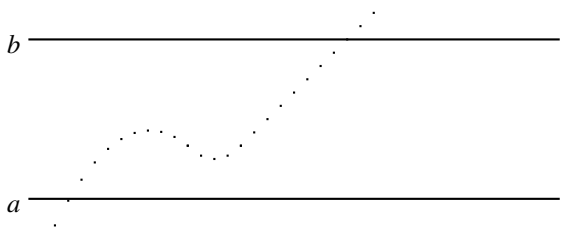
Certainly one of the most amazing things about sub-martingales is the convergence theorem. Recall the earlier sub-martingale convergence theorem. I am going to present the same thing here in this more general setting. Then later, I will present it again in the case of continuous sub-martingales. This which follows is almost identical to the earlier proof.

So why did I even bother with the earlier development? It is because I wanted to make a smooth transition from the idea we usually have of conditional probability where it is probability of something given the value of something else to this more general and much more abstract notion involving σ algebras.

An upcrossing occurs when a sequence goes from a up to b. Thus it crosses the interval, $[a,b]$ in the up direction, hence the name upcrossing. More precisely,

Definition 29.3.3 *Let $\{x_i\}_{i=1}^I$ be any sequence of real numbers, $I \leq \infty$. Define an increasing sequence of integers $\{m_k\}$ as follows. m_1 is the first integer ≥ 1 such that $x_{m_1} \leq a$, m_2 is the first integer larger than m_1 such that $x_{m_2} \geq b$, m_3 is the first integer larger than m_2 such that $x_{m_3} \leq a$, etc. Then each sequence, $\{x_{m_{2k-1}}, \cdots, x_{m_{2k}}\}$, is called an upcrossing of $[a,b]$.*

Here is a picture of an upcrossing.



Proposition 29.3.4 *Let $\{X_i\}_{i=1}^n$ be a finite sequence of real random variables defined on Ω where $(\Omega, \mathscr{F}, P)$ is a probability space. Let $U_{[a,b]}(\omega)$ denote the number of upcrossings of $X_i(\omega)$ of the interval $[a,b]$. Then $U_{[a,b]}$ is a random variable.*

Proof: Let $X_0(\omega) \equiv a+1$, let $Y_0(\omega) \equiv 0$, and let $Y_k(\omega)$ remain 0 for $k = 0, \cdots, l$ until $X_l(\omega) \leq a$. When this happens (if ever), $Y_{l+1}(\omega) \equiv 1$. Then let $Y_i(\omega)$ remain 1 for $i = l+1, \cdots, r$ until $X_r(\omega) \geq b$ when $Y_{r+1}(\omega) \equiv 0$. Let $Y_k(\omega)$ remain 0 for $k \geq r+1$ until $X_k(\omega) \leq a$ when $Y_k(\omega) \equiv 1$ and continue in this way. Thus the upcrossings of $X_i(\omega)$

are identified as unbroken strings of ones for Y_k with a zero at each end, with the possible exception of the last string of ones which may be missing the zero at the upper end and may or may not be an upcrossing.

Note also that Y_0 is measurable because it is identically equal to 0 and that if Y_k is measurable, then Y_{k+1} is measurable because the only change in going from k to $k+1$ is a change from 0 to 1 or from 1 to 0 on a measurable set determined by X_k. In particular,

$$Y_{k+1}^{-1}(1) = ([Y_k = 1] \cap [X_k < b]) \cup ([Y_k = 0] \cap [X_k \leq a])$$

This set is in $\mathscr{F}$ by induction. Of course, $Y_{k+1}^{-1}(0)$ is just the complement of this set. Thus Y_{k+1} is $\mathscr{F}$ measurable since $0,1$ are the only two values. Now let

$$Z_k(\omega) = \begin{cases} 1 \text{ if } Y_k(\omega) = 1 \text{ and } Y_{k+1}(\omega) = 0, \\ 0 \text{ otherwise,} \end{cases}$$

if $k < n$ and

$$Z_n(\omega) = \begin{cases} 1 \text{ if } Y_n(\omega) = 1 \text{ and } X_n(\omega) \geq b, \\ 0 \text{ otherwise.} \end{cases}$$

Thus $Z_k(\omega) = 1$ exactly when an upcrossing has been completed and each Z_i is a random variable.

$$U_{[a,b]}(\omega) = \sum_{k=1}^{n} Z_k(\omega)$$

so $U_{[a,b]}$ is a random variable as claimed. ■

The following corollary collects some key observations found in the above construction.

Corollary 29.3.5 *$U_{[a,b]}(\omega) \leq$ the number of unbroken strings of ones in the sequence $\{Y_k(\omega)\}$, there being at most one unbroken string of ones which produces no upcrossing. Also*

$$Y_i(\omega) = \psi_i\left(\left\{X_j(\omega)\right\}_{j=1}^{i-1}\right), \tag{29.6}$$

where ψ_i is some function of the past values of $X_j(\omega)$.

The following is called the upcrossing lemma.

29.3.1 Upcrossings

Lemma 29.3.6 *(upcrossing lemma) Let $\{(X_i, \mathscr{F}_i)\}_{i=1}^{n}$ be a sub-martingale and let*

$$U_{[a,b]}(\omega)$$

be the number of upcrossings of $[a,b]$. Then

$$E\left(U_{[a,b]}\right) \leq \frac{E(|X_n|) + |a|}{b-a}.$$

Proof: Let $\phi(x) \equiv a + (x-a)^+$ so that ϕ is an increasing convex function always at least as large as a. By Lemma 29.3.2 it follows that $\{(\phi(X_k), \mathscr{F}_k)\}$ is also a sub-martingale.

$$\phi(X_{k+r}) - \phi(X_k) = \sum_{i=k+1}^{k+r} \phi(X_i) - \phi(X_{i-1})$$

$$= \sum_{i=k+1}^{k+r} (\phi(X_i) - \phi(X_{i-1}))\, Y_i + \sum_{i=k+1}^{k+r} (\phi(X_i) - \phi(X_{i-1}))\,(1 - Y_i).$$

Observe that Y_i is $\mathscr{F}_{i-1}$ measurable from its construction in Proposition 29.3.4, Y_i depending only on X_j for $j < i$.

Now let the unbroken strings of ones for $\{Y_i(\omega)\}$ be

$$\{k_1, \cdots, k_1 + r_1\}, \{k_2, \cdots, k_2 + r_2\}, \cdots, \{k_m, \cdots, k_m + r_m\} \tag{29.7}$$

where $m = V(\omega) \equiv$ the number of unbroken strings of ones in the sequence $\{Y_i(\omega)\}$. By Corollary 29.3.5 $V(\omega) \geq U_{[a,b]}(\omega)$.

$$\begin{aligned} &\phi(X_n(\omega)) - \phi(X_1(\omega)) \\ = &\sum_{k=1}^{n} (\phi(X_k(\omega)) - \phi(X_{k-1}(\omega)))\, Y_k(\omega) \\ &+ \sum_{k=1}^{n} (\phi(X_k(\omega)) - \phi(X_{k-1}(\omega)))\,(1 - Y_k(\omega)). \end{aligned}$$

The first sum in the above reduces to summing over the unbroken strings of ones because the terms in which $Y_i(\omega) = 0$ contribute nothing. Therefore, observing that for $x > a, \phi(x) = x$,

$$\begin{aligned} &\phi(X_n(\omega)) - \phi(X_1(\omega)) \geq U_{[a,b]}(\omega)(b-a) + 0 + \\ &\sum_{k=1}^{n} (\phi(X_k(\omega)) - \phi(X_{k-1}(\omega)))\,(1 - Y_k(\omega)) \end{aligned} \tag{29.8}$$

where the zero on the right side results from a string of ones which does not produce an upcrossing. It is here that it is important that $\phi(x) \geq a$. Such a string begins with $\phi(X_k(\omega)) = a$ and results in an expression of the form $\phi(X_{k+m}(\omega)) - \phi(X_k(\omega)) \geq 0$ since $\phi(X_{k+m}(\omega)) \geq a$. If X_k had not been replaced with $\phi(X_k)$, it would have been possible for $\phi(X_{k+m}(\omega))$ to be less than a and the zero in the above could have been a negative number This would have been inconvenient.

Next take the expectation of both sides in 29.8. This results in

$$\begin{aligned} E(\phi(X_n) - \phi(X_1)) &\geq (b-a)E\left(U_{[a,b]}\right) \\ &\quad + E\left(\sum_{k=1}^{n} (\phi(X_k) - \phi(X_{k-1}))\,(1 - Y_k)\right) \\ &\geq (b-a)E\left(U_{[a,b]}\right) \end{aligned}$$

The reason for the last inequality where the term at the end was dropped is

$$\begin{aligned} &E((\phi(X_k) - \phi(X_{k-1}))\,(1 - Y_k)) \\ &= E(E((\phi(X_k) - \phi(X_{k-1}))\,(1 - Y_k)\,|\mathscr{F}_{k-1})) \\ &= E((1 - Y_k)E(\phi(X_k)\,|\mathscr{F}_{k-1}) - (1 - Y_k)E(\phi(X_{k-1})\,|\mathscr{F}_{k-1})) \\ &\geq E((1 - Y_k)(\phi(X_{k-1}) - \phi(X_{k-1}))) = 0. \end{aligned}$$

Recall that Y_k is $\mathscr{F}_{k-1}$ measurable and that $(\phi(X_k), \mathscr{F}_k)$ is a sub-martingale. ■

The reason for this lemma is to prove the amazing sub-martingale convergence theorem.

29.3.2 The Sub-martingale Convergence Theorem

Theorem 29.3.7 *(sub-martingale convergence theorem) Let*

$$\{(X_i, \mathscr{F}_i)\}_{i=1}^{\infty}$$

be a sub-martingale with $K \equiv \sup E(|X_n|) < \infty$. *Then there exists a random variable* X, *such that* $E(|X|) \leq K$ *and*

$$\lim_{n\to\infty} X_n(\omega) = X(\omega) \text{ a.e.}$$

Proof: Let $a, b \in \mathbb{Q}$ and let $a < b$. Let $U_{[a,b]}^n(\omega)$ be the number of upcrossings of $\{X_i(\omega)\}_{i=1}^n$. Then let

$$U_{[a,b]}(\omega) \equiv \lim_{n\to\infty} U_{[a,b]}^n(\omega) = \text{ number of upcrossings of } \{X_i\}.$$

By the upcrossing lemma,

$$E\left(U_{[a,b]}^n\right) \leq \frac{E(|X_n|)+|a|}{b-a} \leq \frac{K+|a|}{b-a}$$

and so by the monotone convergence theorem,

$$E\left(U_{[a,b]}\right) \leq \frac{K+|a|}{b-a} < \infty$$

which shows $U_{[a,b]}(\omega)$ is finite a.e., for all $\omega \notin S_{[a,b]}$ where $P\left(S_{[a,b]}\right) = 0$. Define

$$S \equiv \cup\left\{S_{[a,b]} : a, b \in \mathbb{Q}, a < b\right\}.$$

Then $P(S) = 0$ and if $\omega \notin S$, $\{X_k\}_{k=1}^{\infty}$ has only finitely many upcrossings of every interval having rational endpoints. For such ω it cannot be the case that

$$\limsup_{k\to\infty} X_k(\omega) > \liminf_{k\to\infty} X_k(\omega)$$

because then you could pick rational a, b such that $[a,b]$ is between the limsup and the liminf and there would be infinitely many upcrossings of $[a,b]$. Thus, for $\omega \notin S$,

$$\limsup_{k\to\infty} X_k(\omega) = \liminf_{k\to\infty} X_k(\omega) = \lim_{k\to\infty} X_k(\omega) \equiv X_\infty(\omega).$$

Letting $X_\infty(\omega) \equiv 0$ for $\omega \in S$, Fatou's lemma implies

$$\int_\Omega |X_\infty| \, dP = \int_\Omega \liminf_{n\to\infty} |X_n| \, dP \leq \liminf_{n\to\infty} \int_\Omega |X_n| \, dP \leq K \blacksquare$$

As a simple application, here is an easy proof of a nice theorem about convergence of sums of independent random variables.

Theorem 29.3.8 *Let* $\{X_k\}$ *be a sequence of independent real valued random variables such that* $E(|X_k|) < \infty, E(X_k) = 0$, *and*

$$\sum_{k=1}^{\infty} E\left(X_k^2\right) < \infty.$$

Then $\sum_{k=1}^{\infty} X_k$ *converges a.e.*

Proof: Let $\mathscr{F}_n \equiv \sigma(X_1, \cdots, X_n)$. Consider $S_n \equiv \sum_{k=1}^n X_k$.

$$E(S_{n+1}|\mathscr{F}_n) = S_n + E(X_{n+1}|\mathscr{F}_n).$$

Letting $A \in \mathscr{F}_n$ it follows from independence that

$$\begin{aligned}\int_A E(X_{n+1}|\mathscr{F}_n)\,dP &\equiv \int_A X_{n+1}dP = \int_\Omega \mathscr{X}_A X_{n+1}dP \\ &= P(A)\int_\Omega X_{n+1}dP = 0\end{aligned}$$

and so $E(S_{n+1}|\mathscr{F}_n) = 0$. Therefore, $\{S_n\}$ is a martingale. Now using independence again,

$$E(|S_n|) \le E\left(|S_n^2|\right) = \sum_{k=1}^n E\left(X_k^2\right) \le \sum_{k=1}^\infty E\left(X_k^2\right) < \infty$$

and so $\{S_n\}$ is an L^1 bounded martingale. Therefore, it converges a.e. ■

Corollary 29.3.9 *Let $\{X_k\}$ be a sequence of independent real valued random variables such that $E(|X_k|) < \infty, E(X_k) = m_k$, and*

$$\sum_{k=1}^\infty E\left(|X_k - m_k|^2\right) < \infty.$$

Then $\sum_{k=1}^\infty (X_k - m_k)$ converges a.e.

This can be extended to the case where the random variables have values in a separable Hilbert space. Recall that for $\{e_k\}$ an orthonormal basis and $\sum_{k=1}^\infty |a_k|_H^2 < \infty, \sum_{k=1}^\infty a_k e_k \in H$ and the infinite sum makes sense. Also, for $x \in H, x = \sum_k (x, e_k) e_k$ the convergence in H.

Theorem 29.3.10 *Let $\{X_k\}$ be a sequence of independent H valued random variables where H is a real separable Hilbert space such that $E(|X_k|_H) < \infty, E(X_k) = 0$, and $\sum_{k=1}^\infty E\left(|X_k|_H^2\right) < \infty$. Then $\sum_{k=1}^\infty X_k$ converges a.e.*

Proof: Let $\{e_k\}$ be an orthonormal basis for H. Then $\{(X_n, e_k)_H\}_{n=1}^\infty$ are real valued, independent, and their mean equals 0. Also

$$\sum_{n=1}^\infty E\left(\left|(X_n, e_k)_H^2\right|\right) \le \sum_{n=1}^\infty E\left(|X_n|_H^2\right) < \infty$$

and so from Theorem 29.3.8, the series, $\sum_{n=1}^\infty (X_n, e_k)_H$ converges a.e. Therefore, there exists a set of measure zero such that for ω not in this set, $\sum_n (X_n(\omega), e_k)_H$ converges for each k. For ω not in this exceptional set, define

$$Y_k(\omega) \equiv \sum_{n=1}^\infty (X_n(\omega), e_k)_H$$

Next define $S(\omega) \equiv \sum_{k=1}^\infty Y_k(\omega) e_k$. Of course it is not clear this even makes sense. I need to show $\sum_{k=1}^\infty |Y_k(\omega)|^2 < \infty$. Using the independence of the X_n

$$\begin{aligned}E\left(|Y_k|^2\right) &\le \liminf_{N\to\infty} E\left(\left(\sum_{n=1}^N \sum_{m=1}^N (X_n, e_k)_H (X_m, e_k)_H\right)\right) \\ &= \liminf_{N\to\infty} E\left(\sum_{n=1}^N (X_n, e_k)_H^2\right) = \sum_{n=1}^\infty E\left((X_n, e_k)_H^2\right)\end{aligned}$$

the last from the monotone convergence theorem. Hence from the above,

$$\begin{aligned} E\left(\sum_k |Y_k|^2\right) &= \sum_k E\left(|Y_k|^2\right) \leq \sum_k \sum_n E\left((X_n, e_k)_H^2\right) \\ &= \sum_n E\left(\sum_k (X_n, e_k)^2\right) = \sum_n E\left(|X_n|_H^2\right) < \infty \end{aligned}$$

the last by assumption. Therefore, for ω off a set of measure zero, and for $Y_k(\omega) \equiv \sum_{n=1}^{\infty} (X_n(\omega), e_k)_H$ which exists a.e. by Theorem 29.3.8, $\sum_k |Y_k(\omega)|^2 < \infty$ and so for a.e. ω, $S(\omega) \equiv \sum_{k=1}^{\infty} Y_k(\omega) e_k$ makes sense. Thus for these ω

$$\begin{aligned} S(\omega) &= \sum_l (S(\omega), e_l) e_l = \sum_l Y_l(\omega) e_l \equiv \sum_l \sum_n (X_n(\omega), e_l)_H e_l \\ &= \sum_n \sum_l (X_n(\omega), e_l) e_l = \sum_n X_n(\omega). \blacksquare \end{aligned}$$

Now with this theorem, here is a strong law of large numbers.

Theorem 29.3.11 *Suppose $\{\boldsymbol{X}_k\}$ are independent random variables and*

$$E(|\boldsymbol{X}_k|) < \infty$$

for each k and $E(\boldsymbol{X}_k) = \boldsymbol{m}_k$. Suppose also

$$\sum_{j=1}^{\infty} \frac{1}{j^2} E\left(\left|\boldsymbol{X}_j - \boldsymbol{m}_j\right|^2\right) < \infty. \tag{29.9}$$

Then $\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} (\boldsymbol{X}_j - \boldsymbol{m}_j) = \boldsymbol{0}$ a.e.

Proof: Consider the sum

$$\sum_{j=1}^{\infty} \frac{\boldsymbol{X}_j - \boldsymbol{m}_j}{j}.$$

This sum converges *a.e.* because of 29.9 and Theorem 29.3.10 applied to the random vectors $\left\{\frac{\boldsymbol{X}_j - \boldsymbol{m}_j}{j}\right\}$. Therefore, from Lemma 26.8.4 it follows that for a.e. ω,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} (\boldsymbol{X}_j(\omega) - \boldsymbol{m}_j) = \boldsymbol{0}. \blacksquare$$

The next corollary is often called the strong law of large numbers. It follows immediately from the above theorem.

Corollary 29.3.12 *Suppose $\{\boldsymbol{X}_j\}_{j=1}^{\infty}$ are independent random vectors, $\lambda_{\boldsymbol{X}_i} = \lambda_{\boldsymbol{X}_j}$ for all i, j having mean $\boldsymbol{m}$ and variance equal to*

$$\sigma^2 \equiv \int_{\Omega} \left|\boldsymbol{X}_j - \boldsymbol{m}\right|^2 dP < \infty.$$

Then for a.e. $\omega \in \Omega$, $\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \boldsymbol{X}_j(\omega) = \boldsymbol{m}$

29.3.3 Doob Sub-martingale Estimates

Another very interesting result about sub-martingales is the Doob sub-martingale estimate. First is a technical lemma which is frequently useful in situations where we want to interchange order of integration. I shall likely use this lemma without comment occasionally.

Lemma 29.3.13 *If f is $\mathscr{F}$ measurable and nonnegative then*

$$(\lambda,\omega)\to \mathscr{X}_{[f>\lambda]} \text{ is } \mathscr{F}\times\mathscr{B}(\mathbb{R}) \text{ measurable.}$$

Proof: Let s be a nonnegative simple function, $s(\omega)=\sum_{k=1}^n c_k\mathscr{X}_{E_k}(\omega)$ where we can let the sum be written such that the c_k are strictly increasing in k and these are the positive values of s. Also let $F_k=\cup_{i=k}^n E_i$.

$$\mathscr{X}_{[s>\lambda]}=\sum_{k=1}^n \mathscr{X}_{[c_{k-1},c_k)}(\lambda)\,\mathscr{X}_{F_k}(\omega),\ c_0\equiv 0.$$

which is clearly product measurable. To see that this formula is valid, first consider the case where $\lambda\in[0,c_1)$. Then $\mathscr{X}_{[s>\lambda]}=1$ on F_1 and 0 off F_1. The first term of the right side equals 1 and the others are 0 due to $\mathscr{X}_{[c_{k-1},c_k)}(\lambda)$. Thus the formula holds for such λ. Now suppose $\lambda\in[c_{j-1},c_j)$. Then left side is 1 when $s(\omega)=c_l$ for some $l\geq c_j$. In this case, the right side has exactly one term equal to 1 and it is $\mathscr{X}_{[c_{j-1},c_j)}(\lambda)\,\mathscr{X}_{F_j}(\omega)$. The remaining case is that $\lambda\geq c_n$. In this case, the right side equals 0 and the left side also equals 0 because $s(\omega)$ is never strictly larger than c_n. ■

For arbitrary $f\geq 0$ and measurable, there is an increasing sequence of simple functions s_n converging pointwise to f. Therefore,

$$\lim_{n\to\infty}\mathscr{X}_{[s_n>\lambda]}=\mathscr{X}_{[f>\lambda]}$$

and so $\mathscr{X}_{[f>\lambda]}$ is product measurable. ■

Theorem 29.3.14 *Let $\{(X_i,\mathscr{F}_i)\}_{i=1}^\infty$ be a sub-martingale. Then for $\lambda>0$,*

$$P\left(\left[\max_{1\leq k\leq n}X_k>\lambda\right]\right)\leq\frac{1}{\lambda}\int_\Omega \mathscr{X}_{\left[\max_{1\leq k\leq n}X_k>\lambda\right]}X_n^+dP\leq\frac{1}{\lambda}\int_\Omega X_n^+dP$$

Proof: Let

$$\begin{aligned} A_1 &\equiv [X_1>\lambda], A_2\equiv[X_2>\lambda]\setminus A_1,\\ \cdots, A_k &\equiv [X_k>\lambda]\setminus\left(\cup_{i=1}^{k-1}A_i\right)\cdots\end{aligned}$$

Thus each A_k is $\mathscr{F}_k$ measurable, the A_k are disjoint, and their union equals

$$\left[\max_{1\leq k\leq n}X_k>\lambda\right].$$

Therefore from the definition of a sub-martingale and Jensen's inequality,

$$P\left(\left[\max_{1\leq k\leq n}X_k>\lambda\right]\right)=\sum_{k=1}^n P(A_k)\leq\frac{1}{\lambda}\sum_{k=1}^n\int_{A_k}X_kdP$$

$$\leq \frac{1}{\lambda}\sum_{k=1}^{n}\int_{A_k} E\left(X_n|\mathscr{F}_k\right)dP \leq \frac{1}{\lambda}\sum_{k=1}^{n}\int_{A_k} E\left(X_n|\mathscr{F}_k\right)^+ dP$$

$$\leq \frac{1}{\lambda}\sum_{k=1}^{n}\int_{A_k} E\left(X_n^+|\mathscr{F}_k\right)dP = \frac{1}{\lambda}\sum_{k=1}^{n}\int_{A_k} X_n^+ dP \leq \frac{1}{\lambda}\int_\Omega \mathscr{X}_{\left[\max_{1\leq k\leq n} X_k>\lambda\right]} X_n^+ dP. \blacksquare$$

Now suppose X_k is a martingale with values in W a Banach space. Then from the theorem on conditional expectation, $E\left(\|X_{k+1}\|\,|\mathscr{F}_k\right) \geq \|E\left(X_{k+1}|\mathscr{F}_k\right)\| = \|X_k\|$. Thus $k \to \|X_k\|$ is a sub-martingale and so one gets the following interesting corollary.

Corollary 29.3.15 *Let X_n be a martingale with values in a Banach space W. Then for $\lambda > 0$,*

$$P\left(\left[\max_{1\leq k\leq n}\|X_k\| > \lambda\right]\right) \leq \frac{1}{\lambda}\int_\Omega \mathscr{X}_{\left[\max_{1\leq k\leq n}\|X_k\|>\lambda\right]}\|X_n\|dP \leq \frac{1}{\lambda}\int_\Omega \|X_n\|dP$$

Now suppose X_k is a martingale with values in W a Banach space. For $p>1, k\to\|X_k\|^p$ is a sub-martingale because

$$E\left(\|X_{k+1}\|^p\,|\mathscr{F}_k\right) \geq \left(E\left(\|X_{k+1}\|\,|\mathscr{F}_k\right)\right)^p \geq \|E\left(X_{k+1}|\mathscr{F}_k\right)\|^p = \|X_k\|^p$$

Therefore, from the definition of the Lebesgue integral of a positive function,

$$\int_\Omega \left(\max_{1\leq k\leq n}\|X_k\|\right)^p dP = \int_\Omega \max_{1\leq k\leq n}\|X_k\|^p dP = \int_0^\infty P\left(\left[\max_{1\leq k\leq n}\|X_k\|^p > \lambda\right]\right)d\lambda$$

Change variables $\lambda = \mu^p$ and using the Doob estimate, Theorem 29.3.14,

$$= \int_0^\infty P\left(\left[\max_{1\leq k\leq n}\|X_k\| > \lambda^{1/p}\right]\right)d\lambda = p\int_0^\infty P\left(\left[\max_{1\leq k\leq n}\|X_k\| > \mu\right]\right)\mu^{p-1}d\mu$$

To save on notation, let $X_n^* \equiv \max_{1\leq k\leq n}\|X_k\|$. Then using Lemma 29.3.13 as needed,

$$\leq \int_0^\infty \frac{p\mu^{p-1}}{\mu}\int_\Omega \mathscr{X}_{[X_n^*>\mu]}\|X_n\|dPd\mu = \int_\Omega \|X_n(\omega)\|\int_0^\infty \frac{p\mu^{p-1}}{\mu}\mathscr{X}_{[X_n^*>\mu]}d\mu dP$$

Then $p-1 = p/q$ where $1/p+1/q=1$,

$$\leq \int_\Omega\int_0^{X_n^*} p\mu^{p-2}\|X_n\|d\mu dP = \frac{p}{p-1}\int_\Omega (X_n^*)^{p/q}\|X_n\|dP$$

$$\leq \frac{p}{p-1}\left(\int_\Omega (X_n^*)^p\right)^{1/q}\left(\int_\Omega \|X_n\|^p dP\right)^{1/p}$$

This proves the following version of the above Doob estimate.

Theorem 29.3.16 *Let $n\to X_n$ be a martingale with values in a Banach space W and $X_n \in L^p(\Omega;W), p>1$, then for $X_n^* \equiv \max_{k\leq n}\{\|X_k\|\}$, then*

$$\int_\Omega (X_n^*)^p dP \leq \frac{p}{p-1}\left(\int_\Omega (X_n^*)^p\right)^{1/q}\left(\int_\Omega \|X_n\|^p dP\right)^{1/p}$$

In fact,

$$\left(\int_\Omega (X_n^*)^p dP\right)^{1/p} \leq \frac{p}{p-1}\left(\int_\Omega \|X_n\|^p dP\right)^{1/p}$$

Proof: The second claim is all which remains. Let n be given. Then let

$$A_1 \equiv \{\omega : \|X_1(\omega)\| \geq \|X_k(\omega)\| \text{ for } k \neq 1\}.$$

Let

$$A_2 \equiv \{\omega : \|X_2(\omega)\| \geq \|X_k(\omega)\| k, k \neq 2\} \setminus A_1,$$

etc. Then these sets are disjoint and so

$$X_n^* = \sum_{k=1}^n \mathcal{X}_{A_k} \|X_k\|, \int_\Omega (X_n^*)^p \, dP = \sum_{k=1}^n \int_{A_k} \|X_k\|^p \, dP < \infty$$

and so, we can divide both sides with $\left(\int_\Omega (X_n^*)^p\right)^{1/q}$ to obtain the last claim. ■

Later on I will consider the case of continuous sub-martingales and martingales. In this case, you have to work harder and one way is to use a stopping time. These stopping times are about to be discussed in the next section for discreet processes.

29.4 Optional Sampling and Stopping Times

I think that the optional sampling theorem of Doob is amazing. That is why it gets repeated quite a bit. It is one of those theorems that you read and when you get to the end, having followed the argument, you sit back and feel amazed at what you just went through. You ask yourself if it is really true or whether you made some mistake. At least this is how it affects me.

First it is necessary to define the notion of a stopping time. If you have an increasing sequence of σ algebras $\{\mathcal{F}_n\}$ and a process $\{X_n\}$ such that X_n is $\mathcal{F}_n$ measurable, the idea of a stopping time τ is that τ is measurable and $X_{\min(\tau,n)}$ is a $\mathcal{F}_n$ measurable function. In other words, by stopping with this stopping time, we preserve the $\mathcal{F}_n$ measurability. It is customary to write $n \wedge \tau = \min(n \wedge \tau)$. Thus, we want to have $X_{n\wedge\tau}^{-1}(O) \in \mathcal{F}_n$ where O is an open set in some metric space where X_n has its values and $a \wedge b$ means $\min(a,b)$.

$$X_{n\wedge\tau}^{-1}(O) = [\tau \leq n] \cap \left[\omega : X_{\tau(\omega)}(\omega) \in O\right] \cup [\tau > n] \cap [\omega : X_n(\omega) \in O]$$

Now

$$[\tau \leq n] \cap \left[\omega : X_{\tau(\omega)}(\omega) \in O\right] = \cup_{k=1}^n [\tau = k] \cap [X_k \in O]$$

To have this in $\mathcal{F}_n$, one needs $[\tau = k] \in \mathcal{F}_k$ for each $k \leq n$. That is $[\tau \leq k] \in \mathcal{F}_k$. Now once this is done, $[\tau > n] = [\tau \leq n]^C \in \mathcal{F}_n$ also. This motivates the following definition and shows that the requirement $[\tau \leq n] \in \mathcal{F}_n$ implies that $\omega \to X_{n\wedge\tau(\omega)}(\omega)$ is $\mathcal{F}_n$ measurable when X_n is $\mathcal{F}_n$ measurable and is exactly what is needed for this to happen.

Definition 29.4.1 *Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $\{\mathcal{F}_n\}_{n=1}^\infty$ be an increasing sequence of σ algebras each contained in $\mathcal{F}$. A stopping time is a measurable function τ which maps Ω to $\mathbb{N}$,*

$$\tau^{-1}(A) \in \mathcal{F} \text{ for all } A \in \mathcal{P}(\mathbb{N}),$$

such that for all $n \in \mathbb{N}$,

$$[\tau \leq n] \in \mathcal{F}_n.$$

Note this is equivalent to saying $[\tau = n] \in \mathscr{F}_n$ *because* $[\tau = n] = [\tau \leq n] \setminus [\tau \leq n-1]$. *For* τ *a stopping time define* $\mathscr{F}_\tau$ *as follows.*

$$\mathscr{F}_\tau \equiv \{A \in \mathscr{F} : A \cap [\tau \leq n] \in \mathscr{F}_n \text{ for all } n \in \mathbb{N}\} \tag{29.10}$$

These sets in $\mathscr{F}_\tau$ *are referred to as "prior" to* τ.

Lemma 29.4.2 *The requirement 29.10 is equivalent to saying that* $A \cap [\tau = n] \in \mathscr{F}_n$ *for all* $n \in \mathbb{N}$.

Proof: $[\tau = n] \cap A = [\tau \leq n] \cap A \setminus [\tau \leq n-1] \cap A \in \mathscr{F}_n$ so if 29.10 holds, then $A \cap [\tau = n] \in \mathscr{F}_n$ for all n. Conversely, if $A \cap [\tau = n] \in \mathscr{F}_n$, then $A \cap [\tau \leq n] = \cup_{k \leq n} A \cap [\tau = k]$ which is the union of sets in $\mathscr{F}_n$ since the $\mathscr{F}_k$ are increasing in k. ■

Another good thing to observe is the following lemma.

Lemma 29.4.3 *If* $\sigma \leq \tau$ *and* σ, τ *are two stopping times, then* $\mathscr{F}_\sigma \subseteq \mathscr{F}_\tau$.

Proof: Say $A \in \mathscr{F}_\sigma$ which means that $A \cap [\sigma \leq n] \in \mathscr{F}_n$ for all n. Now consider $A \cap [\tau = n]$. Is this in $\mathscr{F}_n$? Since $\sigma \leq \tau$,

$$A \cap [\tau = n] = \overset{\in \mathscr{F}_n}{[\tau = n]} \cap \cup_{j=1}^{n} \overset{\in \mathscr{F}_n}{A \cap [\sigma = j]} \in \mathscr{F}_n$$

since each $[\sigma = j] \cap A \in \mathscr{F}_n$. ■

Next is a significant observation that a stochastic process $A(n)$ where $A(n)$ is $\mathscr{F}_n$ measurable satisfies $A(\tau) \in \mathscr{F}_\tau$.

Proposition 29.4.4 *Let* $A(k)$ *be* $\mathscr{F}_k$ *measurable for each* k *where* $\mathscr{F}_k$ *is an increasing sequence of* σ *algebras. Then* $A(\tau)$ *is* $\mathscr{F}_\tau$ *measurable if* τ *is a stopping time corresponding to the* $\mathscr{F}_k$. *If* σ, τ *are two stopping times, then so are* $\tau \wedge \sigma$ *and* $\tau \vee \sigma$, *the minimum and maximum of the two stopping times.*

Proof: I need to show that for O an open set $[A(\tau) \in O]$ is $\mathscr{F}_\tau$ measurable. I need to show that $[A(\tau) \in O] \cap [\tau \leq k] \in \mathscr{F}_k$ for each k. It suffices to show that for each j, $[\tau = j] \cap [A(j) \in O] \cap [\tau \leq k] \in \mathscr{F}_k$ for each k. If $j \leq k$, then left side reduces to $[\tau = j] \cap [A(j) \in O]$. However, $A(j)$ is $\mathscr{F}_j$ measurable and so $[A(j) \in O] \in \mathscr{F}_j \subseteq \mathscr{F}_k$ while $[\tau = j] \in \mathscr{F}_j \subseteq \mathscr{F}_k$ so all is well if $j \leq k$. However, if $j > k$, then the expression $[\tau = j] \cap [A(j) \leq r] \cap [\tau \leq k] = \emptyset \in \mathscr{F}_k$ and so it works in this case also. Thus $A(\tau)$ is indeed $\mathscr{F}_\tau$ measurable.

For the last claim, $[\tau \wedge \sigma \leq j] = [\tau \leq j] \cup [\sigma \leq j] \in \mathscr{F}_j$ and $[\tau \vee \sigma \leq j] = [\tau \leq j] \cap [\sigma \leq j] \in \mathscr{F}_j$. ■

Example 29.4.5 *As an example of a stopping time, let* X_n *be* $\mathscr{F}_n$ *measurable where the* $\mathscr{F}_n$ *are increasing* σ *algebras. Let* O *be a Borel set, and let* $\tau(\omega)$ *be the first* n *such that* $X_n(\omega) \in O$. *If* $X_n^{-1}(O) = \emptyset$ *for all* n, *then* $\tau(\omega) \equiv \infty$ *and we consider* $\mathscr{F}_\infty$ *to be* $\mathscr{F}$. *This is an example of a stopping time called the first hitting time. With these discreet processes, it is enough to let* O *be Borel.*

Lemma 29.4.6 *The first hitting time of a Borel set* O *is a stopping time. Also* $\mathscr{F}_\tau$ *is a* σ *algebra.*

Proof: $[\tau \le n] = \cup_{k=1}^{n} [\tau = k] = \cup_{k=1}^{n} [X_k \in O] \cap \left(\cup_{j<k} \left[X_j \in O^C\right]\right)$. Now

$$\overset{\in \mathscr{F}_k}{[X_k \in O]} \cap \left(\cup_{j<k} \overset{\in \mathscr{F}_j}{\left[X_j \in O^C\right]} \right) \in \mathscr{F}_n$$

and so this is indeed a stopping time, being the union of finitely many sets of $\mathscr{F}_n$. As just noted, this means that $X_{n\wedge\tau}$ is $\mathscr{F}_n$ measurable.

For the claim about $\mathscr{F}_\tau$, it is obvious that $\Omega, \emptyset$ are in $\mathscr{F}_\tau$. Suppose $A \in \mathscr{F}_\tau$. Then for k arbitrary, $A^C \cap [\tau \le k] \cup A \cap \overset{\in \mathscr{F}_k}{[\tau \le k]} = \overset{\in \mathscr{F}_k}{[\tau \le k]}$ and so $A^C \cap [\tau \le k] \in \mathscr{F}_k$. It is even more obvious that $\mathscr{F}_\tau$ is closed with respect to countable unions. ■

Of course τ has values i, in a countable well ordered set of numbers, $i \le i+1$. We have the following about the relation with stopping times and conditional expectations.

Lemma 29.4.7 *Let X be in $L^1(\Omega)$. Then*

1. *$\mathscr{F}_\tau \cap [\tau = i] = \mathscr{F}_i \cap [\tau = i]$ and $E(X|\mathscr{F}_\tau) = E(X|\mathscr{F}_i)$ a.e. on the set $[\tau = i]$. Also if $A \in \mathscr{F}_\tau$ or $\mathscr{F}_i$, then $A \cap [\tau = i] \in \mathscr{F}_i \cap \mathscr{F}_\tau$.*

2. *$E(X|\mathscr{F}_\tau) = E(X|\mathscr{F}_i)$ a.e. on the set $[\tau \le i]$.*

Proof: 1.) A typical set in $\mathscr{F}_\tau \cap [\tau = i]$ is $B \equiv A \cap [\tau = i]$ where $A \in \mathscr{F}_\tau$. Thus $A \cap [\tau = i] = B \in \mathscr{F}_i$ so $A \cap [\tau = i] = A \cap [\tau = i] \cap [\tau = i] = B \cap [\tau = i] \in \mathscr{F}_i \cap [\tau = i]$.

A typical set in $\mathscr{F}_i \cap [\tau = i]$ is $A \cap [\tau = i]$ where $A \in \mathscr{F}_i$. Then $A \cap [\tau = i] \cap [\tau = j] \in \mathscr{F}_j$ for all j. If $j \ne i$, you get $\emptyset$ and if $j = i$, you get $A \cap [\tau = i] \in \mathscr{F}_i = \mathscr{F}_j$ so $A \cap [\tau = i] = B \in \mathscr{F}_\tau$ and so $A \cap [\tau = i] \cap [\tau = i] = B \cap [\tau = i] \in \mathscr{F}_\tau \cap [\tau = i]$.

For $A \in \mathscr{F}_\tau$, $A \cap [\tau = i] \in \mathscr{F}_i$ by definition. However, it is also the case, from what was just shown that $A \cap [\tau = i] \in \mathscr{F}_\tau$ because $A \cap [\tau = i] \cap [\tau = j] \in \mathscr{F}_j$ for every j. Also, if $A \in \mathscr{F}_i$, then $A \cap [\tau = i] \in \mathscr{F}_i$ by definition of a stopping time and $A \cap [\tau = i] \cap [\tau = j] \in \mathscr{F}_j$ for every j. Thus if A is either in $\mathscr{F}_\tau$ or $\mathscr{F}_i$, then $[\tau = i] \cap A \in \mathscr{F}_i \cap \mathscr{F}_\tau$.

Now let $A \in \mathscr{F}_\tau$. Then

$$\int_{A\cap[\tau=i]} E(X|\mathscr{F}_i)\, dP = \int_{A\cap[\tau=i]} X dP \equiv \int_{A\cap[\tau=i]} E(X|\mathscr{F}_\tau)\, dP$$

2.) If $A \in \mathscr{F}_\tau$,then $A \cap [\tau = j] \in \mathscr{F}_j$ by definition of $\mathscr{F}_\tau$ and so by definition of conditional expectation,

$$\int_{A\cap[\tau\le i]} E(X|\mathscr{F}_i)\, dP = \sum_{j=1}^{i} \int_{A\cap[\tau=j]} E(X|\mathscr{F}_i)\, dP = \sum_{j=1}^{i} \int_{A\cap[\tau=j]} E(X|\mathscr{F}_j)\, dP$$

$$= \sum_{j=1}^{i} \int_{A\cap[\tau=j]} E(X|\mathscr{F}_\tau)\, dP = \int_{A\cap[\tau\le i]} E(X|\mathscr{F}_\tau)\, dP$$

Since $A \in \mathscr{F}_\tau$ is arbitrary, it follows that $E(X|\mathscr{F}_i) = E(X|\mathscr{F}_\tau)$ a.e. on the set $[\tau \le i]$. ■

29.4.1 Optional Sampling for Martingales

Now it is time for the optional sampling theorem. Suppose $\{X_n\}$ is a martingale. In particular, each $X_n \in L^1(\Omega;W)$ and $E(X_n|\mathscr{F}_k) = X_k$ whenever $k \leq n$. We can assume X_n has values in some separable Banach space. Then $\|X_n\|$ is a sub-martingale because if $k \leq n$, then if $A \in \mathscr{F}_k$,

$$\int_A E(\|X_n\||\mathscr{F}_k)\,dP \geq \int_A \|E(X_n|\mathscr{F}_k)\|\,dP = \int_A \|X_k\|\,dP$$

Now suppose we have two stopping times τ and σ and τ is bounded meaning it has values in $\{1,2,\cdots,n\}$.

The optional sampling theorem says the following. For $M(n)$ a martingale, $\|M(n)\| \in L^1$, then the following holds a.e.

$$M(\sigma\wedge\tau) = E(M(\tau)|\mathscr{F}_\sigma)$$

Furthermore, it all makes sense. First of all, why does it make sense? We need to verify that $M(\tau)$ is integrable.

$$\int \|M(\tau)\| = \sum_{k=1}^n \int_{[\tau=k]} \|M(k)\|\,dP < \infty$$

Similarly, $M(\sigma\wedge\tau)$ is integrable. The reason is that $\sigma\wedge\tau$ is a stopping time which is bounded by n. Thus the above follows with τ replaced with $\sigma\wedge\tau$. Why is $\sigma\wedge\tau$ a stopping time? It is because $[\sigma\wedge\tau \leq k] = [\sigma \leq k]\cup[\tau\leq k] \in \mathscr{F}_k$. It is also clear that $\tau = i \in \mathbb{N}$ will be a stopping time.

Now let $A \in \mathscr{F}_\sigma$. Then using Lemma 29.4.7 as needed,

$$\int_A M(\sigma\wedge\tau) = \sum_{i=1}^n \int_{A\cap[\tau=i]} M(\sigma\wedge i) = \sum_{i=1}^n \sum_{j=1}^\infty \int_{A\cap[\tau=i]\cap[\sigma=j]} M(j\wedge i)$$

If $j \leq i$,

$$M(j\wedge i) = M(j) = E(M(i)|\mathscr{F}_j).$$

If $j > i$,

$$M(j\wedge i) = M(i) = E(M(i)|\mathscr{F}_j).$$

On $[j=\sigma]$, $E(M(i)|\mathscr{F}_j) = E(M(i)|\mathscr{F}_\sigma)$ Thus the last term in the above is

$$= \sum_{i=1}^n \sum_{j=1}^\infty \int_{A\cap[\tau=i]\cap[\sigma=j]} E(M(i)|\mathscr{F}_\sigma) = \sum_{i=1}^n \int_{A\cap[\tau=i]} E(M(i)|\mathscr{F}_\sigma)$$

Now $\mathscr{X}_{A\cap[\tau=i]}M(i) = \mathscr{X}_{A\cap[\tau=i]}M(\tau)$ so

$$= \sum_{i=1}^n \int E\left(\mathscr{X}_{A\cap[\tau=i]}M(i)|\mathscr{F}_\sigma\right) = \sum_{i=1}^n \int E\left(\mathscr{X}_{A\cap[\tau=i]}M(\tau)|\mathscr{F}_\sigma\right)$$

$$= \int E(\mathscr{X}_A M(\tau)|\mathscr{F}_\sigma) = \int_A E(M(\tau)|\mathscr{F}_\sigma)$$

Since A is an arbitrary element of $\mathscr{F}_\sigma$, this shows the optional sampling theorem that $M(\sigma\wedge\tau) = E(M(\tau)|\mathscr{F}_\sigma)$.

Proposition 29.4.8 *Let M be a martingale having values in some separable Banach space. Let τ be a bounded stopping time and let σ be another stopping time. Then everything makes sense in the following formula and*

$$M(\sigma \wedge \tau) \equiv M_{\sigma\wedge\tau} = E(M(\tau)|\mathscr{F}_\sigma) \text{ a.e.}$$

29.4.2 Optional Sampling Theorem for Sub-Martingales

What about the case where $\{X(n)\}$ is a sub-martingale? Shouldn't there be something like the conclusion of Proposition 29.4.8? This requires a very interesting theorem which involves the decomposition of a sub-martingale into a sum. Recall $\{X(k)\}_{k=1}^{\infty}$ is a sub-martingale if

$$E(X(k+1)|\mathscr{F}_k) \geq X(k)$$

where the $\mathscr{F}_k$ are an increasing sequence of σ algebras in the usual way. The following is the very interesting result about writing a sub-martingale as the sum of an increasing process and a martingale.

Lemma 29.4.9 *Let $\{X(k)\}_{k=0}^{\infty}$ be a sub-martingale adapted to the increasing sequence of σ algebras, $\{\mathscr{F}_k\}$. Then there exists a unique increasing process $\{A(k)\}_{k=0}^{\infty}$ such that $A(0)=0$ and $A(k+1)$ is $\mathscr{F}_k$ measurable for all k and a martingale, $\{M(k)\}_{k=0}^{\infty}$ such that*

$$X(k) = A(k) + M(k).$$

Furthermore, for τ a stopping time, $A(\tau)$ is $\mathscr{F}_\tau$ measurable.

Proof: Define $\sum_{k=0}^{-1} \equiv 0$. First consider the uniqueness assertion. Suppose A is a process which does what it is supposed to do.

$$\sum_{k=0}^{n-1} E(X(k+1)-X(k)|\mathscr{F}_k) = \sum_{k=0}^{n-1} E(A(k+1)-A(k)|\mathscr{F}_k)$$
$$+\sum_{k=0}^{n-1} E(M(k+1)-M(k)|\mathscr{F}_k)$$

Then since $\{M(k)\}$ is a martingale,

$$\sum_{k=0}^{n-1} E(X(k+1)-X(k)|\mathscr{F}_k) = \sum_{k=0}^{n-1} A(k+1)-A(k) = A(n)$$

This shows uniqueness and gives a formula for $A(n)$ assuming it exists. It is only a matter of verifying this does work. Define

$$A(n) \equiv \sum_{k=0}^{n-1} E(X(k+1)-X(k)|\mathscr{F}_k),\ A(0)=0.$$

Then A is increasing because from the definition,

$$A(n+1)-A(n) = E(X(n+1)-X(n)|\mathscr{F}_n) \geq 0.$$

Also from the definition above, $A(n)$ is $\mathscr{F}_{n-1}$ measurable, so consider

$$\{X(k)-A(k)\}.$$

Why is this a martingale?

$$\begin{aligned} & E(X(k+1)-A(k+1)|\mathscr{F}_k) = E(X(k+1)|\mathscr{F}_k) - A(k+1) \\ = \; & E(X(k+1)|\mathscr{F}_k) - \sum_{j=0}^{k} E(X(j+1)-X(j)|\mathscr{F}_j) \\ = \; & E(X(k+1)|\mathscr{F}_k) - E(X(k+1)-X(k)|\mathscr{F}_k) \\ & - \sum_{j=0}^{k-1} E(X(j+1)-X(j)|\mathscr{F}_j) \\ = \; & X(k) - \sum_{j=0}^{k-1} E(X(j+1)-X(j)|\mathscr{F}_j) = X(k)-A(k) \end{aligned}$$

Let $M(k) \equiv X(k)-A(k)$. $A(\tau)$ is $\mathscr{F}_\tau$ measurable by Proposition 29.4.4. ■

Note the nonnegative integers could be replaced with any finite set or ordered countable set of numbers with no change in the conclusions of this lemma or the above optional sampling theorem.

Next consider the case of a sub-martingale.

Theorem 29.4.10 *Let $\{X(k)\}$ be a sub-martingale with respect to the increasing sequence of σ algebras, $\{\mathscr{F}_k\}$ and let σ, τ be two stopping times such that τ is bounded. Then $X(\tau)$ defined as*

$$\omega \to X(\tau(\omega))$$

is integrable and

$$X(\sigma \wedge \tau) \leq E(X(\tau)|\mathscr{F}_\sigma).$$

Proof: The claim about $X(\tau)$ being integrable is the same as done earlier. If $\tau \leq l$,

$$E(|X(\tau(\omega))|) = \sum_{i=1}^{l} \int_{[\tau=i]} |X(i)|\, dP < \infty$$

By Lemma 29.4.9 there is a martingale, $\{M(k)\}$ and an increasing process $\{A(k)\}$ such that $A(k+1)$ is $\mathscr{F}_k$ measurable such that

$$X(k) = M(k) + A(k).$$

Then from the fact A is increasing,

$$\begin{aligned} E(X(\tau)|\mathscr{F}_\sigma) &= E(M(\tau)+A(\tau)|\mathscr{F}_\sigma) = M(\tau\wedge\sigma) + E(A(\tau)|\mathscr{F}_\sigma) \\ &\geq M(\tau\wedge\sigma) + E(A(\tau\wedge\sigma)|\mathscr{F}_\sigma) \\ &= M(\tau\wedge\sigma) + A(\tau\wedge\sigma) \equiv X(\tau\wedge\sigma). \end{aligned}$$

because in the above, it follows from Lemma 29.4.9, $A(\tau\wedge\sigma)$ is $\mathscr{F}_{\tau\wedge\sigma}$ measurable and from Lemma 29.4.3, $\mathscr{F}_{\tau\wedge\sigma} \subseteq \mathscr{F}_\sigma$ and so $E(A(\tau\wedge\sigma)|\mathscr{F}_\sigma) = A(\tau\wedge\sigma)$. ■

Say τ is bounded by n and σ is a stopping time. A useful way to remember the above theorem is in the following proposition.

Proposition 29.4.11 *Suppose τ is a bounded stopping time and σ is a stopping time. Then if $\{X(k)\}$ is a sub-martingale, then $X(1), X(\tau\wedge\sigma), X(\tau)$ is also a sub-martingale.*

Proof: Consider $X(1), X(\tau \wedge \sigma), X(\tau)$. Then

$$E(X(\tau)|\mathscr{F}_{\tau\wedge\sigma}) \geq X(\sigma \wedge \tau \wedge \tau) = X(\sigma \wedge \tau).$$

Also $E(X(\sigma\wedge\tau)|\mathscr{F}_1) \geq X(\sigma\wedge\tau\wedge 1) = X(1)$ so this stochastic process is a sub-martingale. ■

This optional sampling theorem gives a convenient way to consider the Doob maximal estimate presented earlier.

Proposition 29.4.12 *Let $\{X(k)\}$ be a real valued sub-martingale, and let $\lambda > 0$. Then for $X_n^* \equiv \max\{X_k : k \leq n\}$ as earlier,*

$$\int_{[X_n^* \leq \lambda]} X(n)^+ dP \geq \lambda P([X_n^* > \lambda]) = \lambda P\left(\left[(X_n^*)^+ > \lambda\right]\right)$$

Proof: Let $\tau = n$ and let σ be the first hitting time of the set (λ, ∞) by $X(k)$. Then $\omega \in [X_n^* > \lambda]$ if and only if for some $k \leq n, X_k > \lambda$ if and only if $\sigma(\omega) = k$ for some $k \leq n$. By Proposition 29.4.11 $X(1), X(\sigma\wedge n), X(n)$ is a sub-martingale and the set of interest $[X_n^* > \lambda]$ is the one where $\sigma < \infty$. Then

$$\begin{aligned} E(X(n)) &\geq E(X(\sigma\wedge n)) = \int_{[\sigma<\infty]} X(\sigma\wedge n)\, dP + \int_{[\sigma=\infty]} X(\sigma\wedge n)\, dP \\ &= \int_{[\sigma<\infty]} X(\sigma\wedge n)\, dP + \int_{[\sigma=\infty]} X(n)\, dP \\ &\geq \lambda P([\sigma<\infty]) + \int_{[\sigma=\infty]} X(n)\, dP \end{aligned}$$

Therefore,

$$E\left(\mathscr{X}_{[X_n^*>\lambda]} X(n)^+\right) \geq E\left(\mathscr{X}_{[X_n^*>\lambda]} X(n)\right) \geq \lambda P([\sigma<\infty]) = \lambda P([X_n^* > \lambda]).\blacksquare$$

Now let $X_{n*} = \min\{X(k) : k \leq n\}$. What about $P([X_{n*} < -\lambda])$? Let σ be the first hitting time for $(-\infty, -\lambda)$ and note that $[X_{n*} < -\lambda]$ consists of the set of ω where $\sigma(\omega) < \infty$. As noted in Proposition 29.4.11, $X(1), X(\sigma\wedge n), X(n)$ is a sub-martingale. Thus

$$\begin{aligned} \int X(1)\, dP &\leq \int_{[\sigma<\infty]} X(\sigma\wedge n)\, dP + \int_{[\sigma=\infty]} X(\sigma\wedge n)\, dP \\ &= \int_{[\sigma<\infty]} X(\sigma\wedge n)\, dP + \int_{[\sigma=\infty]} X(n)\, dP \end{aligned}$$

It follows that

$$\int X(1)\, dP - \int_{[\sigma=\infty]} X(n)\, dP \leq \int_{[\sigma<\infty]} X(\sigma\wedge n)\, dP \leq -\lambda P([\sigma<\infty])$$

and so,

$$\lambda P([\sigma<\infty]) = \lambda P([X_{n*} < \lambda]) \leq \int |X(1)| + |X(n)|\, dP$$

Therefore, we obtain the following theorem which is a maximal estimate for sub-martingales.

Theorem 29.4.13 *Let $\{X(k)\}$ be a real sub-martingale and let $\lambda > 0$ be given. Then*

$$P([\max\{|X_k|, k = 1,...,n\} > \lambda]) \leq \frac{2}{\lambda}\int |X(1)| + |X(n)|\, dP$$

Proof: $[\max\{|X_k|, k = 1,...,n\} > \lambda] \subseteq [X_n^* > \lambda] \cup [X_{n*} < -\lambda]$ and so

$$\begin{aligned} P([\max\{|X_k|, k = 1,...,n\} > \lambda]) &\leq \frac{1}{\lambda}\int X_n^+ dP + \frac{1}{\lambda}\int (|X(1)| + |X(n)|)\, dP \\ &\leq \frac{2}{\lambda}\int |X(1)| + |X(n)|\, dP. \blacksquare \end{aligned}$$

29.5 Reverse Sub-martingale Convergence Theorem

Sub-martingale: $E(X_{n+1}|\mathscr{F}_n) \geq X_n$. Reverse sub-martingale: $E(X_n|\mathscr{F}_{n+1}) \geq X_{n+1}$ and here the $\mathscr{F}_n$ are decreasing.

Definition 29.5.1 *Let $\{X_n\}_{n=0}^{\infty}$ be a sequence of real random variables such that $E(|X_n|) < \infty$ for all n and let $\{\mathscr{F}_n\}$ be a sequence of σ algebras such that $\mathscr{F}_n \supseteq \mathscr{F}_{n+1}$ for all n. Then $\{X_n\}$ is called a reverse sub-martingale if for all n,*

$$E(X_n|\mathscr{F}_{n+1}) \geq X_{n+1}.$$

Note it is just like a sub-martingale only the indices and σ algebras are going the other way. Here is an interesting lemma. This lemma gives uniform integrability for a reverse sub-martingale. The application I have in mind in the next lemma is that $\sup_n E(|X_n|) < \infty$ but it is stated more generally and this condition appears to be obtained for free given the existence of X_∞ in the following lemma.

Lemma 29.5.2 *Suppose for each n, $E(|X_n|) < \infty$, X_n is $\mathscr{F}_n$ measurable, $\mathscr{F}_{n+1} \subseteq \mathscr{F}_n$ for all $n \in \mathbb{N}$, and there exist X_∞ $\mathscr{F}_\infty$ measurable such that $\mathscr{F}_\infty \subseteq \mathscr{F}_n$ for all n and X_0 $\mathscr{F}_0$ measurable such that $\mathscr{F}_0 \supseteq \mathscr{F}_n$ for all n such that for all $n \in \{0,1,\cdots\}$,*

$$E(X_n|\mathscr{F}_{n+1}) \geq X_{n+1},\ E(X_n|\mathscr{F}_\infty) \geq X_\infty,$$

where $E(|X_\infty|) < \infty$. Then $\{X_n : n \in \mathbb{N}\}$ is equi-integrable.

Proof:

$$E(X_{n+1}) \leq E(E(X_n|\mathscr{F}_{n+1})) = E(X_n)$$

Therefore, the sequence $\{E(X_n)\}$ is a decreasing sequence bounded below by $E(X_\infty)$ so it has a limit. I am going to show the functions are equi-integrable. Let k be large enough that

$$\left|E(X_k) - \lim_{m\to\infty} E(X_m)\right| < \varepsilon \tag{29.11}$$

and suppose $n > k$. Then if $\lambda > 0$,

$$\int_{[|X_n|\geq\lambda]} |X_n|\, dP = \int_{[X_n\geq\lambda]} X_n dP + \int_{[X_n\leq-\lambda]} (-X_n)\, dP$$

$$\begin{aligned}
&= \int_{[X_n\geq\lambda]} X_n dP + \int_\Omega (-X_n)\, dP + \int_{[-X_n<\lambda]} X_n dP \\
&\leq \int_{[X_n\geq\lambda]} E(X_k|\mathscr{F}_n)\, dP + \int_\Omega (-X_n)\, dP + \int_{[-X_n<\lambda]} E(X_k|\mathscr{F}_n)\, dP \\
&\leq \int_{[X_n\geq\lambda]} X_k dP + \int_\Omega (-X_k)\, dP + \int_{[X_n>-\lambda]} X_k dP + \varepsilon \\
&= \int_{[X_n\geq\lambda]} X_k dP - \left(\int_\Omega X_k - \int_{[X_n>-\lambda]} X_k dP\right) + \varepsilon
\end{aligned}$$

$$\begin{aligned}
&= \int_{[X_n\geq\lambda]} X_k dP - \left(\int_{[X_n\leq-\lambda]} X_k dP\right) + \varepsilon \\
&= \int_{[X_n\geq\lambda]} X_k dP + \left(\int_{[-X_n\geq\lambda]} (-X_k)\, dP\right) + \varepsilon = \int_{[|X_n|\geq\lambda]} |X_k|\, dP + \varepsilon
\end{aligned}$$

Applying the maximal inequality for sub-martingales, Theorem 29.4.13,

$$P\left(\left[\max\left\{\left|X_j\right| : j = n,\cdots,1\right\} \geq \lambda\right]\right) \leq \frac{1}{\lambda}\left(E(|X_0|) + E(|X_\infty|)\right) \leq \frac{C}{\lambda}$$

and taking sup for all n, $P\left(\left[\sup\left\{\left|X_j\right|\right\} \geq \lambda\right]\right) \leq \frac{C}{\lambda}$. From the above, for $n > k$,

$$\int_{[|X_n|\geq\lambda]} |X_n|\, dP \leq \int_{[|X_n|\geq\lambda]} |X_k|\, dP + \varepsilon,\ P([|x_n| \geq \lambda]) \leq \frac{C}{\lambda}$$

Since the single function X_k is equi-integrable, it follows that for all λ large enough, $\int_{[|X_n|\geq\lambda]} |X_n|\, dP \leq 2\varepsilon$ for all $n > k$. Since there are only finitely many X_j for $j \leq k$, this shows $\{X_n\}$ is equi-integrable. Hence $\{X_n\}$ is uniformly integrable. ■

Note that this also gives X_n bounded in $L^1(\Omega)$ from Proposition 10.9.6 on Page 299. Now with this lemma and the upcrossing lemma it is easy to prove an important convergence theorem.

Theorem 29.5.3 *Let $\{X_n, \mathscr{F}_n\}_{n=0}^\infty$ be a backwards sub-martingale as described above and suppose $\sup_{n\geq 0} E(|X_n|) < \infty$. Then $\{X_n\}$ converges a.e. and in $L^1(\Omega)$ to a function, $X_\infty \in L^1(\Omega)$.*

Proof: By the upcrossing lemma applied to the sub-martingale $\{X_k\}_{k=0}^N$, the number of upcrossings (Downcrossings is probably a better term. They are upcrossings as n gets smaller.) of the interval $[a,b]$ satisfies the inequality

$$E\left(U_{[a,b]}^N\right) \leq \frac{1}{b-a}C$$

Letting $N \to \infty$, it follows the expected number of upcrossings, $E\left(U_{[a,b]}\right)$ is bounded. Therefore, there exists a set of measure 0 N_{ab} such that if $\omega \notin N_{ab}, U_{[a,b]}(\omega) < \infty$. Let $N = \cup\{N_{ab} : a,b \in \mathbb{Q}\}$. Then for $\omega \notin N$,

$$\limsup_{n\to\infty} X_n(\omega) = \liminf_{n\to\infty} X_n(\omega)$$

because if inequality holds, then letting

$$\liminf_{n\to\infty} X_n(\omega) < a < b < \limsup_{n\to\infty} X_n(\omega)$$

it would follow $U_{[a,b]}(\omega) = \infty$, contrary to $\omega \notin N_{ab}$.

Let $X_\infty(\omega) \equiv \lim_{n\to\infty} X_n(\omega)$. Then by Fatou's lemma,

$$\int_\Omega |X_\infty(\omega)|\, dP \le \liminf_{n\to\infty} \int_\Omega |X_n|\, dP < \infty.$$

and so X_∞ is in $L^1(\Omega)$. By the Vitali convergence theorem and Lemma 29.5.2 which shows $\{|X_n|\}$ is uniformly integrable, it follows

$$\lim_{n\to\infty} \int_\Omega |X_\infty(\omega) - X_n(\omega)|\, dP = 0. \blacksquare$$

29.6 Strong Law of Large Numbers

There is a version of the strong law of large numbers which does not depend on the random variables having finite variance. First are some preparatory lemmas. The approach followed here is from Ash [3].

The message of the following lemma, $E(X_k|\sigma(S_n)) = E(X_k|\sigma(S_n, \boldsymbol{Y}))$ makes sense. The expectation of X_k for $k \le n$ given the value of $S_n \equiv \sum_{k=1}^n X_k$ is the same as the expectation of X_k given the value of S_n and $X_{n+1}, \ldots$. It makes intuitive sense because the random variables are independent so knowledge of X_j for $j \ge n+1$ should be irrelevant to the expectation of X_k.

Lemma 29.6.1 *Let $\{X_n\}$ be a sequence of independent random variables such that $E(|X_k|) < \infty$ for all k and let $S_n \equiv \sum_{k=1}^n X_k$. Then for $k \le n$,*

$$E(X_k|\sigma(S_n)) = E(X_k|\sigma(S_n, \boldsymbol{Y})) \text{ a.e.} \tag{29.12}$$

where $\boldsymbol{Y} = (X_{n+1}, X_{n+2}, \cdots) \in \mathbb{R}^{\mathbb{N}} \equiv \prod_{i=1}^\infty \mathbb{R}$. Also for $k \le n$ as above,

$$\sigma(S_n, \boldsymbol{Y}) = \sigma(S_n, S_{n+1}, \cdots).$$

Proof: Note that both X_k and S_n are measurable with respect to $\sigma(X_1, \ldots, X_n)$ and $\sigma(X_1, \ldots, X_n)$ and $\sigma(\boldsymbol{Y})$ are independent. Therefore, by Proposition 29.2.1 29.12 holds.

It only remains to prove the last assertion. For $k > 0$,

$$X_{n+k} = S_{n+k} - S_{n+k-1}$$

Thus

$$\begin{aligned} \sigma(S_n, \boldsymbol{Y}) &= \sigma(S_n, X_{n+1}, \cdots) \\ &= \sigma(S_n, (S_{n+1} - S_n), (S_{n+2} - S_{n+1}), \cdots) \end{aligned}$$

Thus, by induction, each S_{n+k} is measurable with respect to $\sigma(S_n, \boldsymbol{Y})$ and so,

$$\sigma(S_n, S_{n+1}, \cdots) \subseteq \sigma(S_n, \boldsymbol{Y})$$

To get the other inclusion,

$$\sigma(S_n, S_{n+1}, \cdots) = \sigma(S_n, X_{n+1} + S_n, X_{n+2} + X_{n+1} + S_n, \cdots)$$

so by induction, each X_{n+k} and S_n is measurable with respect to $\sigma(S_n, S_{n+1}, \cdots)$ and so $\sigma(S_n, \boldsymbol{Y}) \subseteq \sigma(S_n, S_{n+1}, \cdots)$. $\blacksquare$

Lemma 29.6.2 *Let $\{X_k\}$ be a sequence of independent identically distributed random variables such that $E(|X_k|) < \infty$. Then letting $S_n = \sum_{k=1}^{n} X_k$, it follows that for $k \leq n$*

$$E(X_k|\sigma(S_n, S_{n+1}, \cdots)) = E(X_k|\sigma(S_n)) = \frac{S_n}{n}.$$

Proof: It was shown in Lemma 29.6.1 the first equality holds. It remains to show the second. Letting $A = S_n^{-1}(B)$ where B is Borel, it follows there exists $B' \subseteq \mathbb{R}^n$ a Borel set such that

$$S_n^{-1}(B) = (X_1, \cdots, X_n)^{-1}(B').$$

Then

$$\int_A E(X_k|\sigma(S_n))\,dP = \int_{S_n^{-1}(B)} X_k dP$$

$$= \int_{(X_1,\cdots,X_n)^{-1}(B')} X_k dP = \int_{(X_1,\cdots,X_n)^{-1}(B')} x_k d\lambda_{(X_1,\cdots,X_n)}$$

$$= \int \cdots \int \mathcal{X}_{(X_1,\cdots,X_n)^{-1}(B')}(\boldsymbol{x})\, x_k d\lambda_{X_1} d\lambda_{X_2} \cdots d\lambda_{X_n}$$

$$= \int \cdots \int \mathcal{X}_{(X_1,\cdots,X_n)^{-1}(B')}(\boldsymbol{x})\, x_l d\lambda_{X_1} d\lambda_{X_2} \cdots d\lambda_{X_n}$$

$$= \int_A E(X_l|\sigma(S_n))\,dP$$

and so since $A \in \sigma(S_n)$ is arbitrary,

$$E(X_l|\sigma(S_n)) = E(X_k|\sigma(S_n))$$

for each $k, l \leq n$. Therefore,

$$S_n = E(S_n|\sigma(S_n)) = \sum_{j=1}^{n} E(X_j|\sigma(S_n)) = nE(X_k|\sigma(S_n)) \text{ a.e.}$$

and so

$$E(X_k|\sigma(S_n)) = \frac{S_n}{n} \text{ a.e.}$$

as claimed. ■

With this preparation, here is the strong law of large numbers for identically distributed random variables.

Theorem 29.6.3 *Let $\{X_k\}$ be a sequence of independent identically distributed random variables such that $E(|X_k|) < \infty$ for all k. Since these are identicaly distributed, $E(|X_k|)$ does not depend on k and so the process is bounded. Letting $m = E(X_k)$,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} X_k(\omega) = m \text{ a.e.}$$

and convergence also takes place in $L^1(\Omega)$.

Proof: Consider the reverse sub-martingale $\{E(X_1|\sigma(S_n,S_{n+1},\cdots))\}$. By Theorem 29.5.3, this converges a.e. and in $L^1(\Omega)$ to a random variable X_∞. However, from Lemma 29.6.2, $E(X_1|\sigma(S_n,S_{n+1},\cdots)) = S_n/n$. Therefore, S_n/n converges a.e. and in $L^1(\Omega)$ to X_∞. I need to argue that X_∞ is constant and also that it equals m. For $a \in \mathbb{R}$ let

$$E_a \equiv [X_\infty \geq a]$$

For a small enough, $P(E_a) \neq 0$. Then since E_a is a tail event for the independent random variables, $\{X_k\}$ it follows from the Kolmogorov zero one law, Theorem 26.7.4, that $P(E_a) = 1$. Let $b \equiv \sup\{a : P(E_a) = 1\}$. The sets, E_a are decreasing as a increases. Let $\{a_n\}$ be a strictly increasing sequence converging to b. Then

$$[X_\infty \geq b] = \cap_n [X_\infty \geq a_n]$$

and so $1 = P(E_b) = \lim_{n\to\infty} P(E_{a_n})$. On the other hand, if $c > b$, then $P(E_c) < 1$ and so $P(E_c) = 0$. Hence $P([X = b]) = 1$. It remains to show $b = m$. This is easy because by the L^1 convergence,

$$b = \int_\Omega X_\infty dP = \lim_{n\to\infty} \int_\Omega \frac{S_n}{n} dP = \lim_{n\to\infty} m = m. \blacksquare$$

Chapter 30

Continuous Stochastic Processes

The change here is that the stochastic process will depend on $t \in I$ an interval rather than $n \in \mathbb{N}$. Everything becomes much more technical.

30.1 Fundamental Definitions and Properties

Here E will be a separable Banach space and $\mathscr{B}(E)$ will be the Borel sets of E. Let $(\Omega, \mathscr{F}, P)$ be a probability space and I will be an interval of $\mathbb{R}$. A set of E valued random variables, one for each $t \in I$, $\{X(t) : t \in I\}$ is called a stochastic process. Thus for each t, $X(t)$ is a measurable function of $\omega \in \Omega$. Set $X(t,\omega) \equiv X(t)(\omega)$. Functions $t \to X(t,\omega)$ are called trajectories. Thus there is a trajectory for each $\omega \in \Omega$. A stochastic process, Y is called a version or a modification of a stochastic process X if for all $t \in I$,

$$X(t,\omega) = Y(t,\omega) \text{ a.e. } \omega$$

There are several descriptions of stochastic processes.

1. X is measurable if $X(\cdot,\cdot) : I \times \Omega \to E$ is $B(I) \times \mathscr{F}$ measurable. Note that a stochastic process X is not necessarily measurable.

2. X is stochastically continuous at $t_0 \in I$ means: for all $\varepsilon > 0$ and $\delta > 0$ there exists $\rho > 0$ such that

   $$P([\|X(t) - X(t_0)\| \geq \varepsilon]) \leq \delta \text{ whenever } |t - t_0| < \rho, t \in I.$$

   Note the above condition says that for each $\varepsilon > 0$,

   $$\lim_{t \to t_0} P([\|X(t) - X(t_0)\| \geq \varepsilon]) = 0.$$

3. X is stochastically continuous if it is stochastically continuous at every $t \in I$.

4. X is stochastically uniformly continuous if for every $\varepsilon, \delta > 0$ there exists $\rho > 0$ such that whenever $s, t \in I$ with $|s - t| < \rho$, it follows

   $$P([\|X(t) - X(s)\| \geq \varepsilon]) \leq \delta.$$

5. X is mean square continuous at $t_0 \in I$ if

   $$\lim_{t \to t_0} E\left(\|X(t) - X(t_0)\|^2\right) \equiv \lim_{t \to t_0} \int_\Omega \|X(t)(\omega) - X(t_0)(\omega)\|^2 \, dP = 0.$$

6. X is mean square continuous in I if it is mean square continuous at every point of I.

7. X is continuous with probability 1 or continuous if $t \to X(t,\omega)$ is continuous for all ω outside some set of measure 0.

8. X is Hölder continuous if $t \to X(t,\omega)$ is Hölder continuous for a.e. ω.

Lemma 30.1.1 *A stochastically continuous process on $[a,b] \equiv I$ is uniformly stochastically continuous on $[a,b] \equiv I$.*

Proof: If this is not so, there exists $\varepsilon, \delta > 0$ and points of I, s_n, t_n such that even though $|t_n - s_n| < \frac{1}{n}$,

$$P([\|X(s_n) - X(t_n)\| \geq \varepsilon]) > \delta. \tag{30.1}$$

Taking a subsequence, still denoted by s_n and t_n there exists $t \in I$ such that the above hold and $\lim_{n\to\infty} s_n = \lim_{n\to\infty} t_n = t$. Then

$$\begin{aligned} & P([\|X(s_n) - X(t_n)\| \geq \varepsilon]) \\ \leq\ & P([\|X(s_n) - X(t)\| \geq \varepsilon/2]) + P([\|X(t) - X(t_n)\| \geq \varepsilon/2]). \end{aligned}$$

But the sum of the last two terms converges to 0 as $n \to \infty$ by stochastic continuity of X at t, violating 30.1 for all n large enough. ■

For a stochastically continuous process defined on a closed and bounded interval, there always exists a measurable version. This is significant because then you can do things with product measure and iterated integrals.

Proposition 30.1.2 *Let X be a stochastically continuous process defined on a closed interval, $I \equiv [a,b]$. Then there exists a measurable version of X.*

Proof: By Lemma 30.1.1 X is uniformly stochastically continuous and so there exists a sequence of positive numbers, $\{\rho_n\}$ such that if $|s-t| < \rho_n$, then

$$P\left(\left[\|X(t) - X(s)\| \geq \frac{1}{2^n}\right]\right) \leq \frac{1}{2^n}. \tag{30.2}$$

Then let $\left\{t_0^n, t_1^n, \cdots, t_{m_n}^n\right\}$ be a partition of $[a,b]$ in which $\left|t_i^n - t_{i-1}^n\right| < \rho_n$. Now define X_n as follows:

$$X_n(t) \equiv \sum_{i=1}^{m_n} X\left(t_{i-1}^n\right) \mathcal{X}_{[t_{i-1}^n, t_i^n)}(t),\ X_n(b) \equiv X(b).$$

Then X_n is obviously $B(I) \times \mathscr{F}$ measurable because it is the sum of functions which are. Consider the set A on which $\{X_n(t,\omega)\}$ is a Cauchy sequence. This set is of the form

$$A = \cap_{n=1}^{\infty} \cup_{m=1}^{\infty} \cap_{p,q \geq m} \left[\left\|X_p - X_q\right\| < \frac{1}{n}\right]$$

and so it is a $B(I) \times \mathscr{F}$ measurable set. Now define

$$Y(t,\omega) \equiv \begin{cases} \lim_{n\to\infty} X_n(t,\omega) \text{ if } (t,\omega) \in A \\ 0 \text{ if } (t,\omega) \notin A \end{cases}$$

I claim $Y(t,\omega) = X(t,\omega)$ for a.e. ω. To see this, consider 30.2. From the construction of X_n, it follows that for each t,

$$P\left(\left[\|X_n(t) - X(t)\| \geq \frac{1}{2^n}\right]\right) \leq \frac{1}{2^n} \tag{30.3}$$

Also, for a fixed t, if $X_n(t,\omega)$ fails to converge to $X(t,\omega)$, then ω must be in infinitely many of the sets,

$$B_n \equiv \left[\|X_n(t) - X(t)\| \geq \frac{1}{2^n}\right]$$

which is a set of measure zero by the Borel Cantelli lemma and 30.3. Recall why this is so.

$$P\left(\cap_{k=1}^{\infty}\cup_{n=k}^{\infty}B_n\right)\leq\sum_{n=k}^{\infty}P\left(B_n\right)<\frac{1}{2^{k-1}}$$

Therefore, for each t, $(t,\omega)\in A$ for a.e. ω. Hence $X(t)=Y(t)$ a.e. and so Y is a measurable version of X. ■

One also has the following lemma about extending a process from a dense subset.

Lemma 30.1.3 *Let D be a dense subset of an interval, $I=[0,T]$ and suppose $X:D\to E$ satisfies*

$$\left|\left|X(d)-X\left(d'\right)\right|\right|\leq C\left|d-d'\right|^{\gamma}$$

for all $d',d\in D$. Then X extends uniquely to a continuous Y defined on $[0,T]$ such that

$$\left|\left|Y(t)-Y\left(t'\right)\right|\right|\leq C\left|t-t'\right|^{\gamma}.$$

Proof: Let $t\in I$ and let $d_k\to t$ where $d_k\in D$. Then $\{X(d_k)\}$ is a Cauchy sequence because $||X(d_k)-X(d_m)||\leq C|d_k-d_m|^{\gamma}$. Therefore, $X(d_k)$ converges. The thing it converges to will be called $Y(t)$. Note this is well defined, giving $X(t)$ if $t\in D$. Also, if $d_k\to t$ and $d_k'\to t$, then $\left|\left|X(d_k)-X\left(d_k'\right)\right|\right|\leq C\left|d_k-d_k'\right|^{\gamma}$ and so $X(d_k)$ and $X\left(d_k'\right)$ converge to the same thing. Therefore, it makes sense to define $Y(t)\equiv\lim_{d\to t}X(d)$. It only remains to verify the estimate. But letting $|d-t|$ and $|d'-t'|$ be small enough,

$$\begin{aligned}\left|\left|Y(t)-Y\left(t'\right)\right|\right| &= \left|\left|X(d)-X\left(d'\right)\right|\right|+\varepsilon\\ &\leq C\left|d'-d\right|+\varepsilon\leq C\left|t-t'\right|+2\varepsilon.\end{aligned}$$

Since ε is arbitrary, this proves the existence part of the lemma. Uniqueness follows from observing that $Y(t)$ must equal $\lim_{d\to t}X(d)$. ■

30.2 Kolmogorov Čentsov Continuity Theorem

Lemma 30.2.1 *Let r_j^m denote $j\left(\frac{T}{2^m}\right)$ where $j\in\{0,1,\cdots,2^m\}$. Also let $D_m=\left\{r_j^m\right\}_{j=1}^{2^m}$ and $D=\cup_{m=1}^{\infty}D_m$. Suppose $X(t)$ satisfies*

$$\left\|X\left(r_{j+1}^k\right)-X\left(r_j^k\right)\right\|\leq 2^{-\gamma k}\tag{30.4}$$

for all $k\geq M$. Then if $d,d'\in D_m$ for $m>n\geq M$ such that $|d-d'|\leq T2^{-n}$, then

$$\left\|X\left(d'\right)-X(d)\right\|\leq 2\sum_{j=n+1}^{m}2^{-\gamma j}.\tag{30.5}$$

Also, there exists a constant C depending on M such that for all $d,d'\in D$,

$$\left\|X(d)-X\left(d'\right)\right\|\leq C\left|d-d'\right|^{\gamma}.$$

Proof: Suppose $d'<d$. Suppose first $m=n+1$. Then $d=(k+1)T2^{-(n+1)}$ and $d'=kT2^{-(n+1)}$. Then from 30.4

$$\left\|X\left(d'\right)-X(d)\right\|\leq 2^{-\gamma(n+1)}\leq 2\sum_{j=n+1}^{n+1}2^{-\gamma j}.$$

Suppose the claim 30.5 is true for some $m > n$ and let $d, d' \in D_{m+1}$ with $|d - d'| < T2^{-n}$. If there is no point of D_m between these, then d', d are adjacent points either in D_m or in D_{m+1}. Consequently,

$$\left|\left|X\left(d'\right) - X\left(d\right)\right|\right| \leq 2^{-\gamma m} < 2\sum_{j=n+1}^{m+1} 2^{-\gamma j}.$$

Assume therefore, there exist points of D_m between d' and d. Let $d' \leq d_1' \leq d_1 \leq d$ where d_1, d_1' are in D_m and d_1' is the smallest element of D_m which is at least as large as d' and d_1 is the largest element of D_m which is no larger than d. Then $|d' - d_1'| \leq T2^{-(m+1)}$ and $|d_1 - d| \leq T2^{-(m+1)}$ while all of these points are in D_{m+1} which contains D_m. Therefore, from 30.4 and induction,

$$\begin{aligned} & \left\|X\left(d'\right) - X\left(d\right)\right\| \\ \leq\ & \left\|X\left(d'\right) - X\left(d_1'\right)\right\| + \left\|X\left(d_1'\right) - X\left(d_1\right)\right\| \\ & + \left\|X\left(d_1\right) - X\left(d\right)\right\| \\ \leq\ & 2 \times 2^{-\gamma(m+1)} + 2\sum_{j=n+1}^{m} 2^{-\gamma j} = 2\sum_{j=n+1}^{m+1} 2^{-\gamma j} \\ \leq\ & 2\left(\frac{2^{-\gamma(n+1)}}{1-2^{-\gamma}}\right) = \left(\frac{2T^{-\gamma}}{1-2^{-\gamma}}\right)\left(T2^{-(n+1)}\right)^{\gamma} \end{aligned} \tag{30.6}$$

It follows the above holds for any $d, d' \in D$ such that $|d - d'| \leq T2^{-n}$ because they are both in some D_m for $m > n$.

Consider the last claim. Let $d, d' \in D, |d - d'| \leq T2^{-M}$. Then d, d' are both in some D_m for $m > M$. The number $|d - d'|$ satisfies

$$T2^{-(n+1)} < \left|d - d'\right| \leq T2^{-n}$$

for large enough $n \geq M$. Just pick the first n such that $T2^{-(n+1)} < |d - d'|$. Then from 30.6,

$$\left\|X\left(d'\right) - X\left(d\right)\right\| \leq \left(\frac{2T^{-\gamma}}{1-2^{-\gamma}}\right)\left(T2^{-(n+1)}\right)^{\gamma} \leq \left(\frac{2T^{-\gamma}}{1-2^{-\gamma}}\right)\left(\left|d - d'\right|\right)^{\gamma}$$

Now $[0, T]$ is covered by 2^M intervals of length $T2^{-M}$ and so for any pair $d, d' \in D$,

$$\left\|X\left(d\right) - X\left(d'\right)\right\| \leq C\left|d - d'\right|^{\gamma}$$

where C is a suitable constant depending on 2^M. ■

For $\gamma \leq 1$, you can show, using convexity arguments, that it suffices to have $C = \left(\frac{2T^{-\gamma}}{1-2^{-\gamma}}\right)^{1/\gamma}\left(2^M\right)^{1-\gamma}$. Of course the case where $\gamma > 1$ is not interesting because it would result in X being a constant.

The following is the amazing Kolmogorov Čentsov continuity theorem [32].

Theorem 30.2.2 *Suppose X is a stochastic process on $[0, T]$. Suppose also that there exists a constant, C and positive numbers α, β such that*

$$E\left(\left\|X\left(t\right) - X\left(s\right)\right\|^{\alpha}\right) \leq C\left|t - s\right|^{1+\beta} \tag{30.7}$$

Then there exists a stochastic process Y such that for a.e. $\omega, t \to Y(t)(\omega)$ is Hölder continuous with exponent $\gamma < \frac{\beta}{\alpha}$ and for each t, $P([\|X(t) - Y(t)\| > 0]) = 0$. ($Y$ is a version of X.)

Proof: Let r_j^m denote $j\left(\frac{T}{2^m}\right)$ where $j \in \{0, 1, \cdots, 2^m\}$. Also let $D_m = \left\{r_j^m\right\}_{j=1}^{2^m}$ and $D = \cup_{m=1}^{\infty} D_m$. Consider the set,

$$[\|X(t) - X(s)\| > \delta]$$

By 30.7,

$$\begin{aligned} P([\|X(t) - X(s)\| > \delta])\,\delta^{\alpha} &\leq \int_{[\|X(t)-X(s)\|>\delta]} \|X(t) - X(s)\|^{\alpha}\, dP \\ &\leq C|t-s|^{1+\beta}. \end{aligned} \tag{30.8}$$

Letting $t = r_{j+1}^k$, $s = r_j^k$,and $\delta = 2^{-\gamma k}$ where $\gamma \in \left(0, \frac{\beta}{\alpha}\right)$, this yields

$$P\left(\left[\left\|X\left(r_{j+1}^k\right) - X\left(r_j^k\right)\right\| > 2^{-\gamma k}\right]\right) \leq C2^{\alpha\gamma k}\left(T2^{-k}\right)^{1+\beta}$$

$$= CT^{1+\beta}2^{k(\alpha\gamma-(1+\beta))}$$

There are 2^k of these differences so letting $N_k = \cup_{j=1}^{2^k}\left[\left\|X\left(r_{j+1}^k\right) - X\left(r_j^k\right)\right\| > 2^{-\gamma k}\right]$ it follows

$$P(N_k) \leq C2^{\alpha\gamma k}\left(T2^{-k}\right)^{1+\beta} 2^k = C2^{k(\alpha\gamma-\beta)}T^{1+\beta}.$$

Since $\gamma < \beta/\alpha$, $\sum_{k=1}^{\infty} P(N_k) \leq CT^{1+\beta}\sum_{k=1}^{\infty} 2^{k(\alpha\gamma-\beta)} < \infty$ and so by the Borel Cantelli lemma, Lemma 26.1.2, there exists a set of measure zero N, such that if $\omega \notin N$, then ω is in only finitely many N_k. In other words, for $\omega \notin N$, there exists $M(\omega)$ such that if $k \geq M(\omega)$, then for each j,

$$\left\|X\left(r_{j+1}^k\right)(\omega) - X\left(r_j^k\right)(\omega)\right\| \leq 2^{-\gamma k}. \tag{30.9}$$

It follows from Lemma 30.2.1 that $t \to X(t)(\omega)$ is Holder continuous on D with Holder exponent γ. Note the constant is a measurable function of ω, depending on the number of measurable N_k which contain ω.

By Lemma 30.1.3, one can define $Y(t)(\omega)$ to be the unique function which extends $d \to X(d)(\omega)$ off D for $\omega \notin N$ and let $Y(t)(\omega) = 0$ if $\omega \in N$. Thus by Lemma 30.1.3 $t \to Y(t)(\omega)$ is Holder continuous. Also, $\omega \to Y(t)(\omega)$ is measurable because it is the pointwise limit of measurable functions

$$Y(t)(\omega) = \lim_{d \to t} X(d)(\omega)\,\mathcal{X}_{N^C}(\omega). \tag{30.10}$$

It remains to verify the claim that $Y(t)(\omega) = X(t)(\omega)$ a.e.

$$\mathcal{X}_{[\|Y(t)-X(t)\|>\varepsilon]\cap N^C}(\omega) \leq \liminf_{d \to t} \mathcal{X}_{[\|X(d)-X(t)\|>\varepsilon]\cap N^C}(\omega)$$

because if $\omega \in N$ both sides are 0 and if $\omega \in N^C$ then the above limit in 30.10 holds and so if $||Y(t)(\omega) - X(t)(\omega)|| > \varepsilon$, the same is true of $||X(d)(\omega) - X(t)(\omega)||$ whenever d is close enough to t and so by Fatou's lemma,

$$\begin{aligned} P([\|Y(t) - X(t)\| > \varepsilon]) &= \int \mathcal{X}_{[\|Y(t)-X(t)\|>\varepsilon]\cap N^C}(\omega)\, dP \\ &\leq \int \liminf_{d\to t} \mathcal{X}_{[\|X(d)-X(t)\|>\varepsilon]}(\omega)\, dP \end{aligned}$$

$$\begin{aligned} &\leq \liminf_{d\to t} \int \mathcal{X}_{[\|X(d)-X(t)\|>\varepsilon]}(\omega)\, dP \\ &\leq \liminf_{d\to t} P\left(\left[\|X(d) - X(t)\|^\alpha > \varepsilon^\alpha\right]\right) \end{aligned}$$

$$\begin{aligned} &\leq \liminf_{d\to t} \varepsilon^{-\alpha} \int_{\left[\|X(d)-X(t)\|^\alpha > \varepsilon^\alpha\right]} \|X(d) - X(t)\|^\alpha\, dP \\ &\leq \liminf_{d\to t} \frac{C}{\varepsilon^\alpha} |d-t|^{1+\beta} = 0. \end{aligned}$$

Therefore,

$$\begin{aligned} P([\|Y(t) - X(t)\| > 0]) &= P\left(\cup_{k=1}^\infty \left[\|Y(t) - X(t)\| > \frac{1}{k}\right]\right) \\ \leq \sum_{k=1}^\infty P\left(\left[\|Y(t) - X(t)\| > \frac{1}{k}\right]\right) &= 0. \blacksquare \end{aligned}$$

A few observations are interesting. In the proof, the following inequality was obtained.

$$\begin{aligned} \left\|X\left(d'\right)(\omega) - X(d)(\omega)\right\| &\leq \frac{2}{T^\gamma\left(1-2^{-\gamma}\right)}\left(T2^{-(n+1)}\right)^\gamma \\ &\leq \frac{2}{T^\gamma\left(1-2^{-\gamma}\right)}\left(\left|d-d'\right|\right)^\gamma \end{aligned}$$

which was so for any $d', d \in D$ with $|d' - d| < T2^{-(M(\omega)+1)}$. Thus the Holder continuous version of X will satisfy

$$\|Y(t)(\omega) - Y(s)(\omega)\| \leq \frac{2}{T^\gamma\left(1-2^{-\gamma}\right)}\left(|t-s|\right)^\gamma$$

provided $|t-s| < T2^{-(M(\omega)+1)}$. Does this translate into an inequality of the form

$$\|Y(t)(\omega) - Y(s)(\omega)\| \leq \frac{2}{T^\gamma\left(1-2^{-\gamma}\right)}\left(|t-s|\right)^\gamma$$

for any pair of points $t, s \in [0,T]$? It seems it does not for any $\gamma < 1$ although it does yield

$$\|Y(t)(\omega) - Y(s)(\omega)\| \leq C\left(|t-s|\right)^\gamma$$

where C depends on the number of intervals having length less than $T2^{-(M(\omega)+1)}$ which it takes to cover $[0,T]$. First note that if $\gamma > 1$, then the Holder continuity will imply $t \to$

$Y(t)(\omega)$ is a constant. Therefore, the only case of interest is $\gamma < 1$. Let s,t be any pair of points and let $s = x_0 < \cdots < x_n = t$ where $|x_i - x_{i-1}| < T2^{-(M(\omega)+1)}$. Then

$$\|Y(t)(\omega) - Y(s)(\omega)\| \leq \sum_{i=1}^{n} \|Y(x_i)(\omega) - Y(x_{i-1})(\omega)\|$$

$$\leq \frac{2}{T^{\gamma}(1-2^{-\gamma})} \sum_{i=1}^{n} (|x_i - x_{i-1}|)^{\gamma} \tag{30.11}$$

How does this compare to $(\sum_{i=1}^{n} |x_i - x_{i-1}|)^{\gamma} = |t-s|^{\gamma}$? This last expression is smaller than the right side of 30.11 for any $\gamma < 1$. Thus for $\gamma < 1$, the constant in the conclusion of the theorem depends on both T and $\omega \notin N$.

In the case where $\alpha \geq 1$, here is another proof of this theorem. It is based on the one in the book by Stroock [55]. This one makes the assumption that $\alpha \geq 1$. It isn't for $\alpha > 0$. This version is sufficient for what is done in this book. The Holder estimate is particularly useful.

Theorem 30.2.3 *Suppose X is a stochastic process on $[0,T]$ having values in the Banach space E. Suppose also that there exists a constant C and positive numbers $\alpha, \beta, \alpha \geq 1$, such that*

$$E\left(\|X(t) - X(s)\|^{\alpha}\right) \leq C|t-s|^{1+\beta} \tag{30.12}$$

Then there exists a stochastic process Y such that for a.e. $\omega, t \to Y(t)(\omega)$ is Hölder continuous with exponent $\gamma < \frac{\beta}{\alpha}$ and for each t, $P([\|X(t) - Y(t)\| > 0]) = 0$. ($Y$ is a version of X.) Also

$$E\left(\sup_{0 \leq s < t \leq T} \frac{\|Y(t) - Y(s)\|}{(t-s)^{\gamma}}\right) \leq C$$

where C depends on α, β, T, γ.

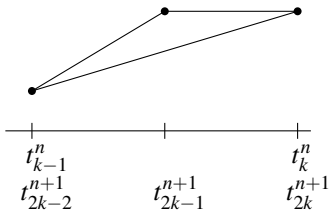
Proof: The proof considers piecewise linear approximations of X which are automatically continuous. These are shown to converge to Y in $L^{\alpha}(\Omega; C([0,T],E))$ so it will follow that Y must be continuous for a.e. ω. Finally, it is shown that Y is a version of X and is Holder continuous. In the proof, I will use C to denote a constant which depends on the quantities γ, α, β, T. Let $\{t_k^n\}_{k=0}^{2^n}$ be a uniform partition of the interval $[0,T]$ so that $t_{k+1}^n - t_k^n = T2^{-n}$. Now let

$$M_n \equiv \max_{k \leq 2^n} \left\|X(t_k^n) - X\left(t_{k-1}^n\right)\right\|$$

Then $M_n^{\alpha} \leq \max_{k \leq 2^n} \left\|X\left(t_k^n\right) - X\left(t_{k-1}^n\right)\right\|^{\alpha} \leq \sum_{k=1}^{2^n} \left\|X\left(t_k^n\right) - X\left(t_{k-1}^n\right)\right\|^{\alpha}$ and so

$$E(M_n^{\alpha}) \leq \sum_{k=1}^{2^n} C\left(T2^{-n}\right)^{1+\beta} = C2^n 2^{-n(1+\beta)} = C2^{-n\beta} \tag{30.13}$$

Next denote by X_n the piecewise linear function which results from the values of X at the points t_k^n. Consider the following picture which illustrates a part of the graphs of X_n and X_{n+1}.



Then

$$\max_{t\in[0,T]} \|X_{n+1}(t) - X_n(t)\| \leq \max_{1\leq k\leq 2^{n+1}} \left\| X\left(t_{2k-1}^{n+1}\right) - \frac{X\left(t_k^n\right) + X\left(t_{k-1}^n\right)}{2} \right\|$$

$$\leq \max_{k\leq 2^{n+1}} \left(\frac{1}{2} \left\| X\left(t_{2k-1}^{n+1}\right) - X\left(t_{2k}^{n+1}\right) \right\| + \frac{1}{2} \left\| X\left(t_{2k-1}^{n+1}\right) - X\left(t_{2k-2}^{n+1}\right) \right\| \right) \leq M_{n+1}$$

Denote by $\|\cdot\|_\infty$ the usual norm in $C([0,T],E)$, $\max_{t\in[0,T]} \|Z(t)\| \equiv \|Z\|_\infty$. Then from what was just established,

$$E\left(\|X_{n+1} - X_n\|_\infty^\alpha\right) = \int_\Omega \|X_{n+1} - X_n\|_\infty^\alpha dP \leq E\left(M_{n+1}^\alpha\right) = C2^{-n\beta}$$

which shows that

$$\|X_{n+1} - X_n\|_{L^\alpha(\Omega;C([0,T],E))} = \left(\int_\Omega \|X_{n+1} - X_n\|_\infty^\alpha dP \right)^{1/\alpha} \leq C\left(2^{(\beta/\alpha)}\right)^{-n}$$

Since $\alpha \geq 1$, we can use the triangle inequality and conclude

$$\|X_m - X_n\|_{L^\alpha(\Omega;C([0,T],E))} \leq$$

$$\sum_{k=n}^{\infty} C\left(2^{(\beta/\alpha)}\right)^{-k} \leq C\frac{\left(2^{(\beta/\alpha)}\right)^{-n}}{1-2^{(-\beta/\alpha)}} = C\left(2^{(\beta/\alpha)}\right)^{-n} \tag{30.14}$$

Thus $\{X_n\}$ is a Cauchy sequence in $L^\alpha(\Omega;C([0,T],E))$ and so it converges to some Y in this space, a subsequence converging pointwise. Then from Fatou's lemma,

$$\|Y - X_n\|_{L^\alpha(\Omega;C([0,T],E))} \leq C\left(2^{(\beta/\alpha)}\right)^{-n}. \tag{30.15}$$

Also, for a.e. $\omega, t \to Y(t)$ is in $C([0,T],E)$. It remains to verify that $Y(t) = X(t)$ a.e.

From the construction, it follows that for any n and $m \geq n$, $Y\left(t_k^n\right) = X_m\left(t_k^n\right) = X\left(t_k^n\right)$. Thus

$$\begin{aligned} \|Y(t) - X(t)\| &\leq \|Y(t) - Y(t_k^n)\| + \|Y(t_k^n) - X(t)\| \\ &= \|Y(t) - Y(t_k^n)\| + \|X(t_k^n) - X(t)\| \end{aligned}$$

Now from the hypotheses of the theorem,

$$P\left(\|X(t_k^n) - X(t)\|^\alpha > \varepsilon\right) \leq \frac{1}{\varepsilon} E\left(\|X(t_k^n) - X(t)\|^\alpha\right) \leq \frac{C}{\varepsilon} |t_k^n - t|^{1+\beta}$$

Thus, there exists a sequence of mesh points $\{s_n\}$ converging to t such that

$$P\left(\|X(s_n)-X(t)\|^{\alpha}>2^{-n}\right)\leq 2^{-n}$$

Then by the Borel Cantelli lemma, there is a set of measure zero N such that for $\omega\notin N$, $\|X(s_n)-X(t)\|^{\alpha}\leq 2^{-n}$ for all n large enough. Then

$$\|Y(t)-X(t)\|\leq\|Y(t)-Y(s_n)\|+\|X(s_n)-X(t)\|$$

which shows that, by continuity of Y, for ω not in an exceptional set of measure zero, $\|Y(t)-X(t)\|=0$.

It remains to verify the assertion about Holder continuity of Y. Let $0\leq s<t\leq T$. Then for some n,

$$2^{-(n+1)}T\leq t-s\leq 2^{-n}T \tag{30.16}$$

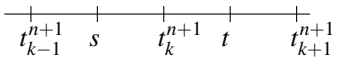
Thus

$$\begin{aligned}\|Y(t)-Y(s)\|&\leq\|Y(t)-X_n(t)\|+\|X_n(t)-X_n(s)\|+\|X_n(s)-Y(s)\|\\&\leq 2\sup_{\tau\in[0,T]}\|Y(\tau)-X_n(\tau)\|+\|X_n(t)-X_n(s)\|\end{aligned} \tag{30.17}$$

Now

$$\frac{\|X_n(t)-X_n(s)\|}{t-s}\leq\frac{\|X_n(t)-X_n(s)\|}{2^{-(n+1)}T}$$

From 30.16 a picture like the following must hold in which $t_{k-1}^{n+1}\leq s<t_k^{n+1}<t\leq t_{k+1}^{n+1}$.



Therefore, from the above, 30.16,

$$\begin{aligned}\frac{\|X_n(t)-X_n(s)\|}{t-s}&\leq\frac{\left\|X\left(t_{k-1}^{n+1}\right)-X\left(t_k^{n+1}\right)\right\|+\left\|X\left(t_{k+1}^{n+1}\right)-X\left(t_k^{n+1}\right)\right\|}{2^{-(n+1)}T}\\&\leq C2^nM_{n+1}\end{aligned}$$

It follows from 30.17,

$$\|Y(t)-Y(s)\|\leq 2\|Y-X_n\|_{\infty}+C2^nM_{n+1}(t-s)$$

Next, letting $\gamma<\beta/\alpha$, and using 30.16,

$$\begin{aligned}\frac{\|Y(t)-Y(s)\|}{(t-s)^{\gamma}}&\leq 2\left(T^{-1}2^{n+1}\right)^{\gamma}\|Y-X_n\|_{\infty}+C2^n\left(2^{-n}\right)^{1-\gamma}M_{n+1}\\&=C2^{n\gamma}\left(\|Y-X_n\|_{\infty}+M_{n+1}\right)\end{aligned}$$

The above holds for any s,t satisfying 30.16. Then

$$\sup\left\{\frac{\|Y(t)-Y(s)\|}{(t-s)^{\gamma}},0\leq s<t\leq T,|t-s|\in\left[2^{-(n+1)}T,2^{-n}T\right]\right\}$$

$$\leq C2^{n\gamma}\left(\|Y-X_n\|_\infty + M_{n+1}\right)$$

Denote by P_n the ordered pairs (s,t) satisfying the above condition that

$$0 \leq s < t \leq T, |t-s| \in \left[2^{-(n+1)}T, 2^{-n}T\right]$$

and also

$$\sup_{(s,t)\in P_n} \frac{\|Y(t)-Y(s)\|}{(t-s)^\gamma} \leq C2^{n\gamma}\left(\|Y-X_n\|_\infty + M_{n+1}\right)$$

Note that the union of the P_n pertains to all (s,t) with $|t-s| \leq T/2$. If $|t-s| > T/2$, then $E\left(\frac{\|Y(t)-Y(s)\|}{|t-s|^\gamma}\right) \leq \left(\frac{2}{T}\right)^\gamma 2\|Y\|_{L^1(\Omega;C([0,T];E))}$ so the desired condition holds and we can ignore this case.

Thus for *a.e.* ω, and for all n,

$$\left(\sup_{(s,t)\in P_n} \frac{\|Y(t)-Y(s)\|}{(t-s)^\gamma}\right)^\alpha \leq C\sum_{k=0}^{\infty} 2^{k\alpha\gamma}\left(\|Y-X_k\|_\infty^\alpha + M_{k+1}^\alpha\right)$$

Note that n is arbitrary. Hence

$$\sup_{0\leq s<t\leq T}\left(\frac{\|Y(t)-Y(s)\|}{(t-s)^\gamma}\right)^\alpha \leq$$

$$\begin{aligned}\sup_n \sup_{(s,t)\in P_n}\left(\frac{\|Y(t)-Y(s)\|}{(t-s)^\gamma}\right)^\alpha &\leq \sup_n\left(\sup_{(s,t)\in P_n}\frac{\|Y(t)-Y(s)\|}{(t-s)^\gamma}\right)^\alpha \\ &\leq \sum_{k=0}^{\infty} C2^{k\alpha\gamma}\left(\|Y-X_k\|_\infty^\alpha + M_{k+1}^\alpha\right)\end{aligned}$$

By continuity of Y, the result on the left is unchanged if the ordered pairs are restricted to lie in $\mathbb{Q}\cap[0,T]\times\mathbb{Q}\cap[0,T]$, a countable set. Thus the left side is measurable. It follows from 30.13 and 30.15 which say

$$\|Y-X_k\|_{L^\alpha(\Omega;C([0,T],E))} \leq C\left(2^{(\beta/\alpha)}\right)^{-k},\ E\left(M_k^\alpha\right) \leq C2^{-k\beta}$$

that

$$E\left(\sup_{0\leq s<t\leq T}\left(\frac{\|Y(t)-Y(s)\|}{(t-s)^\gamma}\right)^\alpha\right) \leq \sum_{k=0}^{\infty} C2^{k\alpha\gamma}2^{-\beta k} \equiv C < \infty$$

because $\alpha\gamma-\beta<0$. By continuity of Y, there are no measurability concerns in taking the above expectation. Note that this implies, since $\alpha \geq 1$,

$$\begin{aligned}E\left(\sup_{0\leq s<t\leq T}\frac{\|Y(t)-Y(s)\|}{(t-s)^\gamma}\right) &\leq \left(E\left(\sup_{0\leq s<t\leq T}\left(\frac{\|Y(t)-Y(s)\|}{(t-s)^\gamma}\right)^\alpha\right)\right)^{1/\alpha} \\ &\leq C^{1/\alpha} \equiv C\end{aligned}$$

Now

$$P\left(\sup_{0\leq s<t\leq T}\left(\frac{\|Y(t)-Y(s)\|}{(t-s)^\gamma}\right)^\alpha > 2^{\alpha k}\right) \leq \frac{1}{2^{\alpha k}}C$$

and so there exists a set of measure zero N such that for $\omega \notin N$,

$$\sup_{0\leq s<t\leq T}\left(\frac{\|Y(t)-Y(s)\|}{(t-s)^{\gamma}}\right)^{\alpha}\leq 2^{\alpha k}$$

for all k large enough. Pick such a k, depending on $\omega \notin N$. Then for any s,t,

$$\frac{\|Y(t)-Y(s)\|}{(t-s)^{\gamma}}\leq 2^{k}$$

and so, this has shown that for $\omega \notin N$, $\|Y(t)-Y(s)\|\leq C(\omega)(t-s)^{\gamma}$ ■

Note that if $X(t)$ is known to be continuous off a set of measure zero, then the piecewise linear approximations converge to $X(t)$ in $C([0,T],E)$ off this set of measure zero. Therefore, it must be that off a set of measure zero, $Y(t)=X(t)$ and so in fact $X(t)$ is Holder continuous off a set of measure zero and the condition on expectation also must hold, that is

$$E\left(\sup_{0\leq s<t\leq T}\frac{\|X(t)-X(s)\|}{(t-s)^{\gamma}}\right)\leq C.$$

30.3 Filtrations

Instead of having a sequence of σ algebras, one can consider an increasing collection of σ algebras indexed by $t\in\mathbb{R}$. This is called a filtration.

Definition 30.3.1 *Let X be a stochastic process defined on an interval, $I=[0,T]$ or $[0,\infty)$. Suppose the probability space, $(\Omega,\mathscr{F},P)$ has an increasing family of σ algebras, $\{\mathscr{F}_t\}$. This is called a filtration. If for arbitrary $t\in I$ the random variable $X(t)$ is $\mathscr{F}_t$ measurable, then X is said to be adapted to the filtration $\{\mathscr{F}_t\}$. Denote by $\mathscr{F}_{t+}$ the intersection of all $\mathscr{F}_s$ for $s>t$. The filtration is normal if*

1. *$\mathscr{F}_0$ contains all $A\in\mathscr{F}$ such that $P(A)=0$*
2. *$\mathscr{F}_t=\mathscr{F}_{t+}$ for all $t\in I$.*

X is called progressively measurable if for every $t\in I$, the mapping

$$(s,\omega)\in[0,t]\times\Omega,\ (s,\omega)\to X(s,\omega)$$

is $B([0,t])\times\mathscr{F}_t$ measurable.

Thus X is progressively measurable means

$$(s,\omega)\to\mathscr{X}_{[0,t]}(s)X(s,\omega)$$

is $B([0,t])\times\mathscr{F}_t$ measurable. As an example of a normal filtration, here is an example.

Example 30.3.2 *For example, you could have a stochastic process, $X(t)$ and you could define*

$$\mathscr{G}_t\equiv\overline{\sigma(X(s):s\leq t)},$$

the completion of the smallest σ algebra such that each $X(s)$ is measurable for all $s \leq t$. This gives an example of a filtration to which $X(t)$ is adapted which satisfies 1. More generally, suppose $X(t)$ is adapted to a filtration, $\mathcal{G}_t$. Define

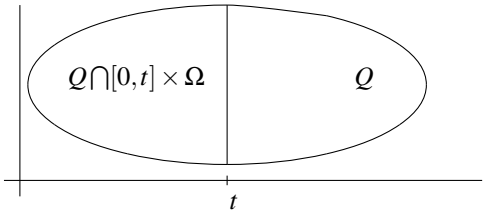
$$\mathcal{F}_t \equiv \cap_{s>t} \mathcal{G}_s$$

Then

$$\mathcal{F}_{t+} \equiv \cap_{s>t} \mathcal{F}_s = \cap_{s>t} \cap_{r>s} \mathcal{G}_r = \cap_{s>t} \mathcal{F}_s \equiv \mathcal{F}_t.$$

and each $X(t)$ is measurable with respect to $\mathcal{F}_t$. Thus there is no harm in assuming a stochastic process adapted to a filtration can be modified so the filtration is normal. Also note that $\mathcal{F}_t$ defined this way will be complete so if $A \in \mathcal{F}_t$ has $P(A) = 0$ and if $B \subseteq A$, then $B \in \mathcal{F}_t$ also. This is because this relation between the sets and the probability of A being zero, holds for this pair of sets when considered as elements of each $\mathcal{G}_s$ for $s > t$. Hence $B \in \mathcal{G}_s$ for each $s > t$ and is therefore one of the sets in $\mathcal{F}_t$.

What is the description of a progressively measurable set?



It means that for Q progressively measurable, $Q \cap [0,t] \times \Omega$ as shown in the above picture is $\mathcal{B}([0,t]) \times \mathcal{F}_t$ measurable. It is like saying a little more descriptively that the function is progressively product measurable.

I shall generally assume the filtration is normal.

Observation 30.3.3 *If X is progressively measurable, then it is adapted. Furthermore the progressively measurable sets, those $E \cap [0,T] \times \Omega$ for which $\mathcal{X}_E$ is progressively measurable form a σ algebra.*

To see why this is, consider X progressively measurable and fix t. Then $(s,\omega) \to X(s,\omega)$ for $(s,\omega) \in [0,t] \times \Omega$ is given to be $\mathcal{B}([0,t]) \times \mathcal{F}_t$ measurable, the ordinary product measure and so fixing any $s \in [0,t]$, it follows the resulting function of ω is $\mathcal{F}_t$ measurable. In particular, this is true upon fixing $s = t$. Thus $\omega \to X(t,\omega)$ is $\mathcal{F}_t$ measurable and so $X(t)$ is adapted.

A set $E \subseteq [0,T] \times \Omega$ is progressively measurable means that $\mathcal{X}_E$ is progressively measurable. That is $\mathcal{X}_E$ restricted to $[0,t] \times \Omega$ is $\mathcal{B}([0,t]) \times \mathcal{F}_t$ measurable. In other words, E is progressively measurable if

$$E \cap ([0,t] \times \Omega) \in \mathcal{B}([0,t]) \times \mathcal{F}_t.$$

If E_i is progressively measurable, does it follow that $E \equiv \cup_{i=1}^{\infty} E_i$ is also progressively measurable? Yes.

$$E \cap ([0,t] \times \Omega) = \cup_{i=1}^{\infty} E_i \cap ([0,t] \times \Omega) \in \mathcal{B}([0,t]) \times \mathcal{F}_t$$

because each set in the union is in $\mathscr{B}([0,t]) \times \mathscr{F}_t$. If E is progressively measurable, is E^C?

$$E^C \cap ([0,t] \times \Omega) \cup \overbrace{(E \cap ([0,t] \times \Omega))}^{\in \mathscr{B}([0,t]) \times \mathscr{F}_t} = \overbrace{[0,t] \times \Omega}^{\in \mathscr{B}([0,t]) \times \mathscr{F}_t}$$

and so $E^C \cap ([0,t] \times \Omega) \in \mathscr{B}([0,t]) \times \mathscr{F}_t$. Thus the progressively measurable sets are a σ algebra.

Another observation of interest is in the following lemma.

Lemma 30.3.4 *Suppose Q is in $\mathscr{B}([0,a]) \times \mathscr{F}_r$. Then if $b \geq a$ and $t \geq r$, then Q is also in $\mathscr{B}([0,b]) \times \mathscr{F}_t$.*

Proof: Consider a measurable rectangle $A \times B$ where $A \in \mathscr{B}([0,a])$ and $B \in \mathscr{F}_r$. Is it true that $A \times B \in \mathscr{B}([0,b]) \times \mathscr{F}_t$? This reduces to the question whether $A \in \mathscr{B}([0,b])$. If A is an interval, it is clear that $A \in \mathscr{B}([0,b])$. Consider the π system of intervals and let $\mathscr{G}$ denote those Borel sets $A \in \mathscr{B}([0,a])$ such that $A \in \mathscr{B}([0,b])$. If $A \in \mathscr{G}$, then $[0,b] \setminus A \in \mathscr{B}([0,b])$ by assumption (the difference of Borel sets is surely Borel). However, this set equals

$$([0,a] \setminus A) \cup (a,b]$$

and so

$$[0,b] = ([0,a] \setminus A) \cup (a,b] \cup A$$

The set on the left is in $\mathscr{B}([0,b])$ and the sets on the right are disjoint and two of them are also in $\mathscr{B}([0,b])$. Therefore, the third, $([0,a] \setminus A)$ is in $\mathscr{B}([0,b])$. It is obvious that $\mathscr{G}$ is closed with respect to countable disjoint unions. Therefore, by Lemma 9.3.2, Dynkin's lemma, $\mathscr{G} \supseteq \sigma(\text{Intervals}) = \mathscr{B}([0,a])$.

Therefore, such a measurable rectangle $A \times B$ where $A \in \mathscr{B}([0,a])$ and $B \in \mathscr{F}_r$ is in $\mathscr{B}([0,b]) \times \mathscr{F}_t$ and in fact it is a measurable rectangle in $\mathscr{B}([0,b]) \times \mathscr{F}_t$. Now let $\mathscr{K}$ denote all these measurable rectangles $A \times B$ where $A \in \mathscr{B}([0,a])$ and $B \in \mathscr{F}_r$. Let $\mathscr{G}$ (new $\mathscr{G}$) denote those sets Q of $\mathscr{B}([0,a]) \times \mathscr{F}_r$ which are in $\mathscr{B}([0,b]) \times \mathscr{F}_t$. Then if $Q \in \mathscr{G}$,

$$Q \cup ([0,a] \times \Omega \setminus Q) \cup (a,b] \times \Omega = [a,b] \times \Omega$$

Then the sets are disjoint and all but $[0,a] \times \Omega \setminus Q$ are in $\mathscr{B}([0,b]) \times \mathscr{F}_t$. Therefore, this one is also in $\mathscr{B}([0,b]) \times \mathscr{F}_t$. If $Q_i \in \mathscr{G}$ and the Q_i are disjoint, then $\cup_i Q_i$ is also in $\mathscr{B}([0,b]) \times \mathscr{F}_t$ and so $\mathscr{G}$ is closed with respect to countable disjoint unions and complements. Hence $\mathscr{G} \supseteq \sigma(\mathscr{K}) = \mathscr{B}([0,a]) \times \mathscr{F}_r$ which shows

$$\mathscr{B}([0,a]) \times \mathscr{F}_r \subseteq \mathscr{B}([0,b]) \times \mathscr{F}_t \quad \blacksquare$$

A significant observation is the following which states that the integral of a progressively measurable function is progressively measurable.

Proposition 30.3.5 *Suppose $X : [0,T] \times \Omega \to E$ where E is a separable Banach space. Also suppose that $X(\cdot,\omega) \in L^1([0,T],E)$ for each ω. Here $\mathscr{F}_t$ is a filtration and with respect to this filtration, X is progressively measurable. Then*

$$(t,\omega) \to \int_0^t X(s,\omega)\, ds$$

is also progressively measurable.

Proof: Suppose $Q \in [0,T] \times \Omega$ is progressively measurable. This means for each t,

$$Q \cap [0,t] \times \Omega \in \mathscr{B}([0,t]) \times \mathscr{F}_t$$

What about $(s,\omega) \in [0,t] \times \Omega$, $(s,\omega) \to \int_0^s \mathscr{X}_Q dr$? Is that function on the right $\mathscr{B}([0,t]) \times \mathscr{F}_t$ measurable? We know that $Q \cap [0,s] \times \Omega$ is $\mathscr{B}([0,s]) \times \mathscr{F}_s$ measurable and hence $\mathscr{B}([0,t]) \times \mathscr{F}_t$ measurable. When you integrate a product measurable function, you do get one which is product measurable. Therefore, this function must be $\mathscr{B}([0,t]) \times \mathscr{F}_t$ measurable. This shows that $(t,\omega) \to \int_0^t \mathscr{X}_Q(s,\omega)\, ds$ is progressively measurable. Here is a claim which was just used.

Claim: If Q is $\mathscr{B}([0,t]) \times \mathscr{F}_t$ measurable, then $(s,\omega) \to \int_0^s \mathscr{X}_Q dr$ is also $\mathscr{B}([0,t]) \times \mathscr{F}_t$ measurable.

Proof of claim: First consider $A \times B$ where $A \in \mathscr{B}([0,t])$ and $B \in \mathscr{F}_t$. Then

$$\int_0^s \mathscr{X}_{A\times B} dr = \int_0^s \mathscr{X}_A \mathscr{X}_B dr = \mathscr{X}_B(\omega) \int_0^s \mathscr{X}_A(s)\, dr$$

This is clearly $\mathscr{B}([0,t]) \times \mathscr{F}_t$ measurable. It is the product of a continuous function of s with the indicator function of a set in $\mathscr{F}_t$. Now let

$$\mathscr{G} \equiv \left\{ Q \in \mathscr{B}([0,t]) \times \mathscr{F}_t : (s,\omega) \to \int_0^s \mathscr{X}_Q(r,\omega)\, dr \text{ is } \mathscr{B}([0,t]) \times \mathscr{F}_t \text{ measurable} \right\}$$

Then it was just shown that $\mathscr{G}$ contains the measurable rectangles. It is also clear that $\mathscr{G}$ is closed with respect to countable disjoint unions and complements. Therefore, $\mathscr{G} \supseteq \sigma(\mathscr{K}_t) = \mathscr{B}([0,t]) \times \mathscr{F}_t$ where $\mathscr{K}_t$ denotes the measurable rectangles $A \times B$ where $B \in \mathscr{F}_t$ and $A \in \mathscr{B}([0,t]) = \mathscr{B}([0,T]) \cap [0,t]$. This proves the claim.

Thus if Q is progressively measurable, $(s,\omega) \to \int_0^s \mathscr{X}_Q(r,\omega)\, dr \equiv f(s,\omega)$ is progressively measurable because for $(s,\omega) \in [0,t] \times \Omega$, $(s,\omega) \to f(s,\omega)$ is $\mathscr{B}([0,t]) \times \mathscr{F}_t$ measurable. This is what was to be proved in this special case.

Now consider the conclusion of the proposition. By considering the positive and negative parts of $\phi(X)$ for $\phi \in E'$, and using Pettis theorem, it suffices to consider the case where $X \geq 0$. Then there exists an increasing sequence of progressively measurable simple functions $\{X_n\}$ converging pointwise to X. From what was just shown,

$$(t,\omega) \to \int_0^t X_n ds$$

is progressively measurable. Hence, by the monotone convergence theorem, $(t,\omega) \to \int_0^t X ds$ is also progressively measurable. ■

What else can you do to something which is progressively measurable and obtain something which is progressively measurable? It turns out that shifts in time can preserve progressive measurability. Let $\mathscr{F}_t$ be a filtration on $[0,T]$ and extend the filtration to be equal to $\mathscr{F}_0$ and $\mathscr{F}_T$ for $t < 0$ and $t > T$, respectively. Recall the following definition of progressively measurable sets.

Definition 30.3.6 *Denote by $\mathscr{P}$ those sets Q in $\mathscr{F}_T \times \mathscr{B}([0,T])$ such that for $t \in [-\infty, T]$*

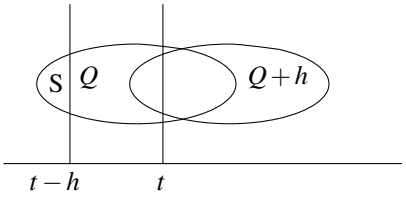
$$\Omega \times (-\infty, t] \cap Q \in \mathscr{F}_t \times \mathscr{B}((-\infty, t]).$$

Lemma 30.3.7 *Define $Q+h$ as*

$$Q+h \equiv \{(t+h,\omega) : (t,\omega) \in Q\}.$$

Then if $Q \in \mathscr{P}$, it follows that $Q+h \in \mathscr{P}$.

Proof: This is most easily seen through the use of the following diagram. In this diagram, Q is in $\mathscr{P}$ so it is progressively measurable.



By definition, S in the picture is $\mathscr{B}((-\infty,t-h]) \times \mathscr{F}_{t-h}$ measurable. Hence $S+h \equiv Q+h \cap \Omega \times (-\infty,t]$ is $\mathscr{B}((-\infty,t]) \times \mathscr{F}_{t-h}$ measurable. To see this, note that if $B \times A \in \mathscr{B}((-\infty,t-h]) \times \mathscr{F}_{t-h}$, then translating it by h gives a set in $\mathscr{B}((-\infty,t]) \times \mathscr{F}_{t-h}$. Then if $\mathscr{G}$ consists of sets S in $\mathscr{B}((-\infty,t-h]) \times \mathscr{F}_{t-h}$ for which $S+h$ is in $\mathscr{B}((-\infty,t]) \times \mathscr{F}_{t-h}$, $\mathscr{G}$ is closed with respect to countable disjoint unions and complements. Thus, $\mathscr{G}$ equals $\mathscr{B}((-\infty,t-h]) \times \mathscr{F}_{t-h}$. In particular, it contains the set S just described. ■

Now for $h > 0$,

$$\tau_h f(t) \equiv \begin{cases} f(t-h) \text{ if } t \geq h, \\ 0 \text{ if } t < h. \end{cases}.$$

Lemma 30.3.8 *Let $Q \in \mathscr{P}$. Then $\tau_h \mathscr{X}_Q$ is $\mathscr{P}$ measurable.*

Proof: If $\tau_h \mathscr{X}_Q(t,\omega) = 1$, then you need to have $(t-h,\omega) \in Q$ and so $(t,\omega) \in Q+h$. Thus

$$\tau_h \mathscr{X}_Q = \mathscr{X}_{Q+h},$$

which is $\mathscr{P}$ measurable since $Q \in \mathscr{P}$. In general,

$$\tau_h \mathscr{X}_Q = \mathscr{X}_{[h,T]\times\Omega} \mathscr{X}_{Q+h},$$

which is $\mathscr{P}$ measurable. ■

This lemma implies the following.

Lemma 30.3.9 *Let $f(t,\omega)$ have values in a separable Banach space and suppose f is $\mathscr{P}$ measurable. Then $\tau_h f$ is $\mathscr{P}$ measurable.*

Proof: Taking values in a separable Banach space and being $\mathscr{P}$ measurable, f is the pointwise limit of $\mathscr{P}$ measurable simple functions. If s_n is one of these, then from the above lemmas, $\tau_h s_n$ is $\mathscr{P}$ measurable. Then, letting $n \to \infty$, it follows that $\tau_h f$ is $\mathscr{P}$ measurable. ■

The following is similar to Proposition 30.1.2. It shows that under pretty weak conditions, an adapted process has a progressively measurable adapted version.

Proposition 30.3.10 *Let X be a stochastically continuous adapted process for a normal filtration defined on a closed interval, $I \equiv [0,T]$. Then X has a progressively measurable adapted version.*

Proof: By Lemma 30.1.1 X is uniformly stochastically continuous and so there exists a sequence of positive numbers, $\{\rho_n\}$ such that if $|s-t|<\rho_n$, then

$$P\left(\left[\|X(t)-X(s)\|\geq\frac{1}{2^n}\right]\right)\leq\frac{1}{2^n}. \tag{30.18}$$

Then let $\left\{t_0^n,t_1^n,\cdots,t_{m_n}^n\right\}$ be a partition of $[0,T]$ in which $\left|t_i^n-t_{i-1}^n\right|<\rho_n$. Now define X_n as follows:

$$\begin{aligned} X_n(t)(\omega) &\equiv \sum_{i=1}^{m_n} X\left(t_{i-1}^n\right)(\omega)\,\mathcal{X}_{[t_{i-1}^n,t_i^n)}(t) \\ X_n(T) &\equiv X(T). \end{aligned}$$

Then $(s,\omega)\to X_n(s,\omega)$ for $(s,\omega)\in[0,t]\times\Omega$ is obviously $B([0,t])\times\mathscr{F}_t$ measurable. Consider the set, A on which $\{X_n(t,\omega)\}$ is a Cauchy sequence. This set is of the form

$$A=\cap_{n=1}^{\infty}\cup_{m=1}^{\infty}\cap_{p,q\geq m}\left[\|X_p-X_q\|<\frac{1}{n}\right]$$

and so it is a $B(I)\times\mathscr{F}$ measurable set and $A\cap[0,t]\times\Omega$ is $B([0,t])\times\mathscr{F}_t$ measurable for each $t\leq T$ because each X_q in the above has the property that its restriction to $[0,t]\times\Omega$ is $B([0,t])\times\mathscr{F}_t$ measurable. Now define

$$Y(t,\omega)\equiv\begin{cases}\lim_{n\to\infty}X_n(t,\omega) \text{ if } (t,\omega)\in A\\ 0 \text{ if } (t,\omega)\notin A\end{cases}$$

I claim that for each t, $Y(t,\omega)=X(t,\omega)$ for a.e. ω. To see this, consider 30.18. From the construction of X_n, it follows that for each t,

$$P\left(\left[\|X_n(t)-X(t)\|\geq\frac{1}{2^n}\right]\right)\leq\frac{1}{2^n}$$

Also, for a fixed t, if $X_n(t,\omega)$ fails to converge to $X(t,\omega)$, then ω must be in infinitely many of the sets,

$$B_n\equiv\left[\|X_n(t)-X(t)\|\geq\frac{1}{2^n}\right]$$

which is a set of measure zero by the Borel Cantelli lemma. Recall why this is so.

$$P(\cap_{k=1}^{\infty}\cup_{n=k}^{\infty}B_n)\leq\sum_{n=k}^{\infty}P(B_n)<\frac{1}{2^{k-1}}$$

Therefore, for each t, $(t,\omega)\in A$ for a.e. ω. Hence $X(t)=Y(t)$ a.e. and so Y is a measurable version of X. Y is adapted because the filtration is normal and hence $\mathscr{F}_t$ contains all sets of measure zero. Therefore, $Y(t)$ differs from $X(t)$ on a set which is $\mathscr{F}_t$ measurable. ■

There is a more specialized situation in which the measurability of a stochastic process automatically implies it is adapted. Furthermore, this can be defined easily in terms of a π system of sets.

Definition 30.3.11 *Let $\mathscr{F}_t$ be a filtration on $(\Omega,\mathscr{F},P)$ and denote by $\mathscr{P}_\infty$ the smallest σ algebra of sets of $[0,\infty)\times\Omega$ containing the sets*

$$(s,t]\times F, F\in\mathscr{F}_s \quad \{0\}\times F,\ F\in\mathscr{F}_0.$$

This is called the predictable σ algebra. and the sets in this σ algebra are called the predictable sets. Denote by $\mathscr{P}_T$ the intersection of $\mathscr{P}_\infty$ to $[0,T]\times\Omega$. A stochastic process X which maps either $[0,T]\times\Omega$ or $[0,\infty)\times\Omega$ to E, a separable real Banach space is called predictable if for every Borel set $A\in\mathscr{B}(E)$, it follows $X^{-1}(A)\in\mathscr{P}_T$ or $\mathscr{P}_\infty$.

This is a lot like product measure except one of the σ algebras is changing.

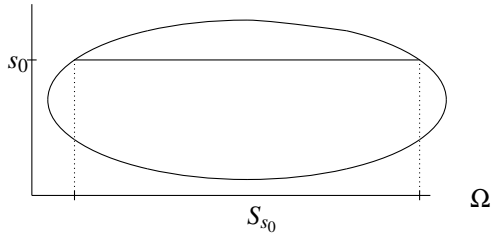
Proposition 30.3.12 *Let $\mathscr{F}_t$ be a filtration as above and let X be a predictable stochastic process. Then X is $\mathscr{F}_t$ adapted.*

Proof: Let $s_0>0$ and define

$$\mathscr{G}_{s_0}\equiv\left\{S\in\mathscr{P}_\infty : S_{s_0}\in\mathscr{F}_{s_0}\right\}$$

where

$$S_{s_0}\equiv\{\omega\in\Omega : (s_0,\omega)\in S\}.$$



It is clear $\mathscr{G}_{s_0}$ is a σ algebra. The next step is to show $\mathscr{G}_{s_0}$ contains the sets

$$(s,t]\times F, F\in\mathscr{F}_s \tag{30.19}$$

and

$$\{0\}\times F,\ F\in\mathscr{F}_0. \tag{30.20}$$

It is clear $\{0\}\times F$ is contained in $\mathscr{G}_{s_0}$ because $(\{0\}\times F)_{s_0}=\emptyset\in\mathscr{F}_{s_0}$. Similarly, if $s\geq s_0$ or if $s,t<s_0$ then $((s,t]\times F)_{s_0}=\emptyset\in\mathscr{F}_{s_0}$. The only case left is for $s<s_0$ and $t\geq s_0$. In this case, letting $A_s\in\mathscr{F}_s$, $((s,t]\times A_s)_{s_0}=A_s\in\mathscr{F}_s\subseteq\mathscr{F}_{s_0}$. Therefore, $\mathscr{G}_{s_0}$ contains all the sets of the form given in 30.19 and 30.20 and so since $\mathscr{P}_\infty$ is the smallest σ algebra containing these sets, it follows $\mathscr{P}_\infty=\mathscr{G}_{s_0}$. The case where $s_0=0$ is entirely similar but shorter.

Therefore, if X is predictable, letting $A\in\mathscr{B}(E)$, $X^{-1}(A)\in\mathscr{P}_\infty$ or $\mathscr{P}_T$ and so

$$\left(X^{-1}(A)\right)_s\equiv\{\omega\in\Omega : X(s,\omega)\in A\}=X(s)^{-1}(A)\in\mathscr{F}_s$$

showing $X(t)$ is $\mathscr{F}_t$ adapted. ■

Another way to see this is to recall the progressively measurable functions are adapted. Then show the predictable sets are progressively measurable.

Proposition 30.3.13 *Let $\mathscr{P}$ denote the predictable σ algebra and let $\mathscr{R}$ denote the progressively measurable σ algebra. Then $\mathscr{P}\subseteq\mathscr{R}$.*

Proof: Let $\mathscr{G}$ denote those sets of $\mathscr{P}$ such that they are also in $\mathscr{R}$. Then $\mathscr{G}$ clearly contains the π system of sets $\{0\}\times A, A\in\mathscr{F}_0$, and $(s,t]\times A, A\in\mathscr{F}_s$. Furthermore, $\mathscr{G}$ is closed with respect to countable disjoint unions and complements. It follows $\mathscr{G}$ contains the σ algebra generated by this π systems which is $\mathscr{P}$. ■

Proposition 30.3.14 *Let $X(t)$ be a stochastic process having values in E a complete metric space and let it be $\mathscr{F}_t$ adapted and left continuous. Then it is predictable. Also, if $X(t)$ is stochastically continuous and adapted on $[0,T]$, then it has a predictable version.*

Proof: Define $I_{m,k} \equiv ((k-1)2^{-m}T, k2^{-m}T]$ if $k \geq 1$ and $I_{m,0} = \{0\}$ if $k = 1$. Then define

$$\begin{aligned} X_m(t) &\equiv \sum_{k=1}^{2^m} X\left(T(k-1)2^{-m}\right) \mathscr{X}_{((k-1)2^{-m}T, k2^{-m}T]}(t) \\ &\quad + X(0)\,\mathscr{X}_{[0,0]}(t) \end{aligned}$$

Here the sum means that $X_m(t)$ has value $X\left(T(k-1)2^{-m}\right)$ on the interval

$$((k-1)2^{-m}T, k2^{-m}T].$$

Thus X_m is predictable because each term in the sum is. Thus

$$\begin{aligned} X_m^{-1}(U) &= \cup_{k=1}^{2^m} \left(X\left(T(k-1)2^{-m}\right)\mathscr{X}_{((k-1)2^{-m}T, k2^{-m}T]}\right)^{-1}(U) \\ &= \cup_{k=1}^{2^m} ((k-1)2^{-m}T, k2^{-m}T] \times \left(X\left(T(k-1)2^{-m}\right)\right)^{-1}(U), \end{aligned}$$

a finite union of predictable sets. Since X is left continuous,

$$X(t,\omega) = \lim_{m\to\infty} X_m(t,\omega)$$

and so X is predictable.

Next consider the other claim. Since X is stochastically continuous on $[0,T]$, it is uniformly stochastically continuous on this interval by Lemma 30.1.1. Therefore, there exists a sequence of partitions of $[0,T]$, the m^{th} being

$$0 = t_{m,0} < t_{m,1} < \cdots < t_{m,n(m)} = T$$

such that for X_m defined as above, then for each t

$$P\left(\left[d(X_m(t), X(t)) \geq 2^{-m}\right]\right) \leq 2^{-m} \tag{30.21}$$

Then as above, X_m is predictable. Let A denote those points of $\mathscr{P}_T$ at which $X_m(t,\omega)$ converges. Thus A is a predictable set because it is just the set where $X_m(t,\omega)$ is a Cauchy sequence. Now define the predictable function Y

$$Y(t,\omega) \equiv \begin{cases} \lim_{m\to\infty} X_m(t,\omega) \text{ if } (t,\omega) \in A \\ 0 \text{ if } (t,\omega) \notin A \end{cases}$$

From 30.21 it follows from the Borel Cantelli lemma that for fixed t, the set of ω which are in infinitely many of the sets,

$$\left[d(X_m(t), X(t)) \geq 2^{-m}\right]$$

has measure zero. Therefore, for each t, there exists a set of measure zero, $N(t)$ such that for $\omega \notin N(t)$ and all m large enough

$$d(X_m(t,\omega), X(t,\omega)) < 2^{-m}$$

Hence for $\omega \notin N(t)$, $(t,\omega) \in A$ and so $X_m(t,\omega) \to Y(t,\omega)$ which shows

$$d(Y(t,\omega), X(t,\omega)) = 0 \text{ if } \omega \notin N(t).$$

The predictable version of $X(t)$ is $Y(t)$. ■

Here is a summary of what has been shown above.

adapted and left continuous
$\Downarrow$
predictable
$\Downarrow$
progressively measurable
$\Downarrow$
adapted

Also

stochastically continuous and adapted $\implies$ progressively measurable version

30.4 Martingales and Sub-Martingales

This was done earlier for discreet martingales. The idea here is to consider indiscreet (What a word to use for a martingale!) ones.

Definition 30.4.1 *Let X be a stochastic process defined on an interval I with values in a separable Banach space, E. It is called integrable if $E(\|X(t)\|) < \infty$ for each $t \in I$. Also let $\mathscr{F}_t$ be a filtration. An integrable and adapted stochastic process X is called a martingale if for $s \leq t$*

$$E(X(t)|\mathscr{F}_s) = X(s) \text{ } P \text{ a.e. } \omega.$$

Recalling the definition of conditional expectation, this says that for $F \in \mathscr{F}_s$

$$\int_F X(t)\,dP = \int_F E(X(t)|\mathscr{F}_s)\,dP = \int_F X(s)\,dP$$

for all $F \in \mathscr{F}_s$. A real valued stochastic process is called a sub-martingale if whenever $s \leq t$,

$$E(X(t)|\mathscr{F}_s) \geq X(s) \text{ } a.e.$$

and a supermartingale if

$$E(X(t)|\mathscr{F}_s) \leq X(s) \text{ } a.e.$$

Example 30.4.2 *Let $\mathscr{F}_t$ be a filtration and let Z be in $L^1(\Omega, \mathscr{F}_T, P)$. Then let $X(t) \equiv E(Z|\mathscr{F}_t)$.*

This works because for $s < t$, $E(X(t)|\mathscr{F}_s) \equiv E(E(Z|\mathscr{F}_t)|\mathscr{F}_s) = E(Z|\mathscr{F}_s) \equiv X(s)$.

Proposition 30.4.3 *The following statements hold for a stochastic process defined on $[0,T] \times \Omega$ having values in a real separable Banach space, E.*

1. *If $X(t)$ is a martingale then $\|X(t)\|, t \in [0,T]$ is a sub-martingale.*

2. *If g is an increasing convex function from $[0,\infty)$ to $[0,\infty)$ and*

$$E(g(\|X(t)\|)) < \infty$$

for all $t \in [0,T]$ then $g(\|X(t)\|), t \in [0,T]$ is a sub-martingale.

Proof: Let $s \leq t$. Then from properties of conditional expecation and Theorem 24.12.1 on Page 729,

$$\begin{aligned} \|X(s)\| &= \|E(X(s)-X(t)|\mathscr{F}_s)+E(X(t)|\mathscr{F}_s)\| \\ &\leq \overbrace{\|E(X(s)-X(t)|\mathscr{F}_s)\|}^{=0 \text{ a.e.}} + \|E(X(t)|\mathscr{F}_s)\| \leq \|E(X(t)|\mathscr{F}_s)\| \\ &\leq E(\|X(t)\||\mathscr{F}_s) \end{aligned}$$

Consider the second claim. Recall Jensen's inequality for sub-martingales, Theorem 29.1.7 on Page 812. From the first part

$$\|X(s)\| \leq E(\|X(t)\||\mathscr{F}_s) \text{ a.e.}$$

and so from Jensen's inequality,

$$g(\|X(s)\|) \leq g(E(\|X(t)\||\mathscr{F}_s)) \leq E(g(\|X(t)\|)|\mathscr{F}_s) \text{ a.e.},$$

showing that $g(\|X(t)\|)$ is also a sub-martingale. ■

30.5 Some Maximal Estimates

Martingales and sub-martingales have some very interesting maximal estimates. I will present some of these here. The proofs are fairly general. For convenience, assume each $\mathscr{F}_t$ contains the sets of measure zero from $\mathscr{F}$. This is so that it suffices to assume $t \to X(t)(\omega)$ is right continuous off some set of measure zero. If it were right continuous for each ω, then it wouldn't matter. Actually, in this book, I will mainly be interested in continuous processes. It is also possible to show that for real valued processes, one can get a right continuous version but this will not be used.

Lemma 30.5.1 *Let $\{\mathscr{F}_t\}$ be a filtration and let $\{X(t)\}$ be a nonnegative valued sub-martingale for $t \in [S,T]$. Then for $\lambda > 0$ and any $p \geq 1$, if, for each t, A_t is a $\mathscr{F}_t$ measurable subset of $[X(t) > \lambda]$, then*

$$P(A_t) \leq \frac{1}{\lambda^p}\int_{A_t} X(T)^p \, dP.$$

Proof: From Jensen's inequality,

$$\begin{aligned} \lambda^p P(A_t) &\leq \int_{A_t} X(t)^p \, dP \leq \int_{A_t} E(X(T)|\mathscr{F}_t)^p \, dP \\ &\leq \int_{A_t} E(X(T)^p|\mathscr{F}_t) \, dP = \int_{A_t} X(T)^p \, dP \blacksquare \end{aligned}$$

The following theorem is the main result.

Theorem 30.5.2 *Let $\{\mathscr{F}_t\}$ be a filtration and let $\{X(t)\}$ be a nonnegative valued right continuous*[1] *sub-martingale for $t \in [S,T]$. Then for all $\lambda > 0$ and $p \geq 1$, for*

$$X^* \equiv \sup_{t\in[S,T]} X(t),$$

$$P([X^* > \lambda]) \leq \frac{1}{\lambda^p}\int_\Omega \mathscr{X}_{[X^*>\lambda]} X(T)^p \, dP$$

In the case that $p > 1$, it is also true that

$$E((X^*)^p) \leq \left(\frac{p}{p-1}\right) E(X(T)^p)^{1/p} (E((X^*)^p))^{1/p'}$$

Also there are no measurability issues related to the above $\sup_{t\in[S,T]} X(t) \equiv X^*$. If $X(t) \in L^p(\Omega)$ for each t, then

$$E((X^*)^p)^{1/p} \leq \left(\frac{p}{p-1}\right) E(X(T)^p)^{1/p}$$

Thus X^* is also in $L^p(\Omega)$.

Proof: Let $S \leq t_0^m < t_1^m < \cdots < t_{N_m}^m = T$ where $t_{j+1}^m - t_j^m = (T-S)2^{-m}$. First consider $m = 1$.

$$A_{t_0^1} \equiv \left\{\omega \in \Omega : X\left(t_0^1\right)(\omega) > \lambda\right\}, \; A_{t_1^1} \equiv \left\{\omega \in \Omega : X\left(t_1^1\right)(\omega) > \lambda\right\} \setminus A_{t_0^1}$$

$$A_{t_2^1} \equiv \left\{\omega \in \Omega : X\left(t_2^1\right)(\omega) > \lambda\right\} \setminus \left(A_{t_0^1} \cup A_{t_0^1}\right).$$

Do this type of construction for $m = 2,3,4,\cdots$ yielding disjoint sets, $\left\{A_{t_j^m}\right\}_{j=0}^{2^m}$ whose union equals

$$\cup_{t\in D_m}[X(t) > \lambda]$$

where $D_m = \left\{t_j^m\right\}_{j=0}^{2^m}$. Thus $D_m \subseteq D_{m+1}$. Then also, $D \equiv \cup_{m=1}^\infty D_m$ is dense and countable. From Lemma 30.5.1,

$$\begin{aligned} P(\cup_{t\in D_m}[X(t) > \lambda]) &= P\left(\left[\sup_{t\in D_m} X(t) > \lambda\right]\right) = \sum_{j=0}^{2^m} P\left(A_{t_j^m}\right) \\ &\leq \frac{1}{\lambda^p}\sum_{j=0}^{2^m}\int_{A_{t_j^m}} \mathscr{X}_{\left[\sup_{t\in D_m} X(t)>\lambda\right]} X(T)^p \, dP \qquad (30.22) \end{aligned}$$

$$\leq \frac{1}{\lambda^p}\int_\Omega \mathscr{X}_{\left[\sup_{t\in D_m} X(t)>\lambda\right]} X(T)^p \, dP \leq \frac{1}{\lambda^p}\int_\Omega \mathscr{X}_{\left[\sup_{t\in D} X(t)>\lambda\right]} X(T)^p \, dP.$$

Let $m \to \infty$ in the above to obtain

$$P(\cup_{t\in D}[X(t) > \lambda]) = P\left(\left[\sup_{t\in D} X(t) > \lambda\right]\right) \leq \frac{1}{\lambda^p}\int_\Omega \mathscr{X}_{\left[\sup_{t\in D} X(t)>\lambda\right]} X(T)^p \, dP. \quad (30.23)$$

[1] $t \to X(t)(\omega)$ is continuous from the right for a.e. ω.

From now on, we begin using the assumption that for a.e. $\omega \in \Omega$, $t \to X(t)(\omega)$ is right continuous. Then with this assumption of right continuity, the following claim holds.

$$\sup_{t\in[S,T]} X(t) \equiv X^* = \sup_{t\in D} X(t)$$

which verifies that X^* is measurable. Then from 30.23,

$$P([X^* > \lambda]) = P\left(\left[\sup_{t\in D} X(t) > \lambda\right]\right)$$

$$\leq \frac{1}{\lambda^p}\int_\Omega \mathcal{X}_{[\sup_{t\in D} X(t)>\lambda]} X(T)^p\, dP = \frac{1}{\lambda^p}\int_\Omega \mathcal{X}_{[X^*>\lambda]} X(T)^p\, dP$$

Now consider the other inequality. Using the distribution function technique and the above estimate obtained in the first part, and earlier facts about the distribution function,

$$E((X^*)^p) = \int_0^\infty p\alpha^{p-1} P([X^* > \alpha])\, d\alpha$$

Then using Lemma 29.3.13 to justify interchange in order of integration,

$$\begin{aligned} &\leq \int_0^\infty p\alpha^{p-1}\frac{1}{\alpha}\int_\Omega \mathcal{X}_{[X^*>\alpha]} X(T)\, dP d\alpha = p\int_\Omega \int_0^{X^*} \alpha^{p-2} d\alpha X(T)\, dP \\ &= \frac{p}{p-1}\int_\Omega (X^*)^{p-1} X(T)\, dP \leq \frac{p}{p-1}\left(\int_\Omega (X^*)^p\right)^{1/p'}\left(\int_\Omega X(T)^p\right)^{1/p} \\ &= \frac{p}{p-1} E(X(T)^p)^{1/p} E((X^*)^p)^{1/p'}. \end{aligned} \tag{30.24}$$

Now assume $X(t) \in L^p(\Omega)$. Returning to 30.22, and letting X_n^* be $\sup_{t\in D_n} X(t)$, this says that

$$P([X_n^* > \lambda]) \leq \frac{1}{\lambda^p}\int_\Omega \mathcal{X}_{[X_n^*>\lambda]} X(T)^p\, dP$$

Then X_n^* achieves its maximum at one of finitely many values for t on a suitable subset of Ω. Thus it makes sense to write $\int_\Omega (X_n^*)^p\, dP$. Now repeat the argument. This yields

$$E((X_n^*)^p) \leq \frac{p}{p-1} E(X(T)^p)^{1/p} E((X_n^*)^p)^{1/p'}$$

Dividing by $E((X_n^*)^p)^{1/p'}$, one obtains

$$(E((X_n^*)^p))^{1/p} \leq \frac{p}{p-1} E(X(T)^p)^{1/p}$$

Now let $n \to \infty$ and use the monotone convergence theorem. ■

If you assumed $t \to X(t)$ is lower semi-continuous instead of right continuous, it appears the above argument would also work.

With Theorem 30.5.2, here is an important maximal estimate for martingales having values in E, a real separable Banach space. In the following, either $t \to X(t)(\omega)$ is right continuous for all ω or for a.e. ω each $\mathcal{F}_t$ contains the sets of measure zero.

Theorem 30.5.3 *Let $X(t)$ for $t \in I = [0,T]$ be an E valued right continuous martingale with respect to a filtration $\mathscr{F}_t$. Then for $p \geq 1$,*

$$P\left(\left[\sup_{t\in I}\|X(t)\| > \lambda\right]\right) \leq \frac{1}{\lambda^p} E\left(\|X(T)\|^p\right). \tag{30.25}$$

If $p > 1$,

$$E\left(\left(\sup_{t\in[S,T]}\|X(t)\|\right)^p\right) \leq \left(\frac{p}{p-1}\right) E\left(\|X(T)\|^p\right)^{1/p} E\left(\left(\sup_{t\in[S,T]}\|X(t)\|\right)^p\right)^{1/p'} \tag{30.26}$$

If, in addition, each $X(t) \in L^p(\Omega)$ for each t, then

$$E\left(\left(\sup_{t\in[S,T]}\|X(t)\|\right)^p\right)^{1/p} \leq \left(\frac{p}{p-1}\right) E\left(\|X(T)\|^p\right)^{1/p} \tag{30.27}$$

Proof: By Proposition 30.4.3 $\|X(t)\|, t \in I$ is a sub-martingale and so from Theorem 30.5.2, it follows 30.25 and 30.26 hold. 30.27 also holds from Theorem 30.5.2. You just apply that theorem to the sub-martingale $Z(t) \equiv \|X(t)\|$ and let $Z^*(t) = \sup_{s\in[S,T]}\|X(s)\|$. ■

Chapter 31

Optional Sampling Theorems

As with discreet martingales, there is a notion of stopping time and optional sampling theorems. These are considered by approximating with discreet stopping times. It is like the case of the integral where one uses step functions or simple functions to approximate a given function.

31.1 Review of Discreet Stopping Times

First it is necessary to define the notion of a stopping time. The following definition was discussed earlier in the context of discreet processes.

Definition 31.1.1 *Let $(\Omega,\mathscr{F},P)$ be a probability space and let $\{\mathscr{F}_n\}_{n=1}^{\infty}$ be an increasing sequence of σ algebras each contained in $\mathscr{F}$, called a discrete filtration. A stopping time is a measurable function, τ which maps Ω to $\mathbb{N}$,*

$$\tau^{-1}(A) \in \mathscr{F} \text{ for all } A \in \mathscr{P}(\mathbb{N}),$$

such that for all $n \in \mathbb{N}$,

$$[\tau \leq n] \in \mathscr{F}_n.$$

Note this is equivalent to saying

$$[\tau = n] \in \mathscr{F}_n$$

because

$$[\tau = n] = [\tau \leq n] \setminus [\tau \leq n-1].$$

For τ a stopping time define $\mathscr{F}_\tau$ as follows.

$$\mathscr{F}_\tau \equiv \{A \in \mathscr{F} : A \cap [\tau \leq n] \in \mathscr{F}_n \text{ for all } n \in \mathbb{N}\}$$

These sets in $\mathscr{F}_\tau$ are referred to as "prior" to τ.

It is clear that $\mathscr{F}_\tau$ is a σ algebra.
The most important example of a stopping time is the first hitting time.

Example 31.1.2 *The first hitting time of an adapted process $X(n)$ of a Borel set G is a stopping time. This is defined as*

$$\tau \equiv \min\{k : X(k) \in G\}$$

To see this, note that

$$[\tau = n] = \cap_{k<n}\left[X(k) \in G^C\right] \cap [X(n) \in G] \in \mathscr{F}_n.$$

This led to the following proposition. It was Proposition 29.4.4.

Proposition 31.1.3 *For τ a stopping time, $\mathscr{F}_\tau$ is a σ algebra and if $Y(k)$ is $\mathscr{F}_k$ measurable for all k, $Y(k)$ having values in a separable Banach space E, then*

$$\omega \to Y(\tau(\omega))$$

is $\mathscr{F}_\tau$ measurable.

To see this,

$$(Y\circ\tau)^{-1}(G)\cap[\tau\leq n]=\cup_k\left(\overbrace{Y(k)^{-1}(G)}^{\in\mathscr{F}_k}\right)\cap[\tau=k]\cap[\tau\leq n]$$

The term in the union is $\emptyset$ if $k>n$ and so the whole thing reduces to

$$\cup_{k=1}^{n}\left(\overbrace{Y(k)^{-1}(G)}^{\in\mathscr{F}_k}\right)\cap[\tau=k]\ \in\mathscr{F}_n$$

showing that $(Y\circ\tau)^{-1}(G)\in\mathscr{F}_\tau$.

The following lemma contains the fundamental properties of stopping times for discrete filtrations. It was Lemma 29.4.7.

Lemma 31.1.4 *In the situation of Definition 31.1.1,*

1. $\mathscr{F}_\tau\cap[\tau=i]=\mathscr{F}_i\cap[\tau=i]$ *and* $E(X|\mathscr{F}_\tau)=E(X|\mathscr{F}_i)$ *a.e. on the set* $[\tau=i]$. *Also if* $A\in\mathscr{F}_\tau$ *or* $\mathscr{F}_i$, *then* $A\cap[\tau=i]\in\mathscr{F}_i\cap\mathscr{F}_\tau$.
2. $E(X|\mathscr{F}_\tau)=E(X|\mathscr{F}_i)$ *a.e. on the set* $[\tau\leq i]$.
3. *Also, if* $\sigma\leq\tau$, *then* $\mathscr{F}_\sigma\subseteq\mathscr{F}_\tau$

Proof: The first two are in the above mentioned lemma. The first part of 1. comes fairly quickly from the definition. The next part of 1. about the conditional expectations is essentially because one can regard $\mathscr{F}_\tau\cap[\tau=i]$ and $\mathscr{F}_i\cap[\tau=i]$ as two equal σ algebras contained in $[\tau=i]$ and so the two conditional expectations are the same on $[\tau=i]$. The third part of 1. also follows from the definition. Then 2. is clearly true from 1. applied to $[\tau=j]$ for $j\leq i$.

Say $A\in\mathscr{F}_\tau$. Then for $j\leq i,[\tau=j]\in\mathscr{F}_\tau$ because $[\tau=j]\cap[\tau\leq k]\in\mathscr{F}_k$ for each k. Thus

$$\begin{aligned}\int_{A\cap[\tau=j]}XdP &= \int_{A\cap[\tau=j]}E(X|\mathscr{F}_\tau)\,dP=\int_{\substack{A\cap[\tau=j]\\ \in\mathscr{F}_j}}E(X|\mathscr{F}_j)\,dP\\ &= \int_{A\cap[\tau=j]}E(X|\mathscr{F}_i)\,dP\end{aligned}$$

Since A is arbitrary, $E(X|\mathscr{F}_\tau)=E(X|\mathscr{F}_i)$.

Now consider 3. If $A\in\mathscr{F}_\sigma$, this means $A\cap[\sigma\leq i]\in\mathscr{F}_i$ or equivalently, $A\cap[\sigma=i]\in\mathscr{F}_i$ for all i. Take such an A. Then $A\cap[\tau=n]=\cup_{i=1}^{n}A\cap[\sigma=i]\in\mathscr{F}_n$ and so $\mathscr{F}_\sigma\subseteq\mathscr{F}_\tau$. ■

The assertion that

$$E(Y|\mathscr{F}_\tau)=E(Y|\mathscr{F}_k)\ \text{a.e.}$$

on $[\tau=k]$ and that a function g which is $\mathscr{F}_\tau$ or $\mathscr{F}_k$ measurable when restricted to $[\tau=k]$ is $\mathscr{G}$ measurable for

$$\mathscr{G}=[\tau=k]\cap\mathscr{F}_\tau=[\tau=k]\cap\mathscr{F}_k$$

is the main result in the above lemma and this fact leads to the amazing Doob optional sampling theorem below. Also note that if $Y(k)$ is any process defined on the positive integers k, then by definition, $Y(k)(\omega)=Y(\tau(\omega))(\omega)$ on the set $[\tau=k]$ because τ is constant on this set.

31.2 Review of Doob Optional Sampling Theorem

With this lemma, here is a major theorem, the optional sampling theorem of Doob. This one is for martingales having values in a Banach space. To begin with, consider the case of a martingale defined on a countable set. This was discussed earlier but it is the sort of thing that seems to me should be repeated because it is so amazing.

Theorem 31.2.1 *Let $\{M(k)\}$ be a martingale having values in E a separable real Banach space with respect to the increasing sequence of σ algebras, $\{\mathscr{F}_k\}$ and let σ,τ be two stopping times such that τ is bounded. Then $M(\tau)$ defined as $\omega \to M(\tau(\omega))$ is integrable and*

$$M(\sigma\wedge\tau) = E(M(\tau)|\mathscr{F}_\sigma).$$

Proof: By Proposition 31.1.3 $M(\tau)$ is $\mathscr{F}_\tau$ measurable.

Next note that since τ is bounded by some l,

$$\int_\Omega \|M(\tau(\omega))\|\,dP \le \sum_{i=1}^{l}\int_{[\tau=i]}\|M(i)\|\,dP < \infty.$$

This proves the first assertion and makes possible the consideration of conditional expectation.

$(E(M(l)|\mathscr{F}_\tau) = M(\tau))$ Let $l \ge \tau$ as described above. Then for $k \le l$, by Lemma 31.1.4,

$$\mathscr{F}_k \cap [\tau = k] = \mathscr{F}_\tau \cap [\tau = k] \equiv \mathscr{G}$$

implying that if g is either $\mathscr{F}_k$ measurable or $\mathscr{F}_\tau$ measurable, then its restriction to $[\tau=k]$ is $\mathscr{G}$ measurable and so if $A \in \mathscr{F}_\tau \cap [\tau=k]$ then

$$\begin{aligned}\int_A E(M(l)|\mathscr{F}_\tau)\,dP &\equiv \int_A M(l)\,dP = \int_A E(M(l)|\mathscr{F}_k)\,dP\\ &= \int_A M(k)\,dP = \int_A M(\tau)\,dP \text{ (on } A, \tau = k)\end{aligned}$$

Therefore, since A was arbitrary, $E(M(l)|\mathscr{F}_\tau) = M(\tau)$ a.e. on $[\tau=k]$ for every $k \le l$. It follows $E(M(l)|\mathscr{F}_\tau) = M(\tau)$ a.e. since it is true on each $[\tau=k]$ for all $k \le l$.

$(M(\sigma\wedge\tau) = E(M(\tau)|\mathscr{F}_\sigma))$ Now consider $E(M(\tau)|\mathscr{F}_\sigma)$ on the set $[\sigma=i]\cap[\tau=j]$. By Lemma 31.1.4, on this set,

$$E(M(\tau)|\mathscr{F}_\sigma) = E(M(\tau)|\mathscr{F}_i) = E(E(M(l)|\mathscr{F}_\tau)|\mathscr{F}_i) = E(E(M(l)|\mathscr{F}_j)|\mathscr{F}_i)$$

If $j \le i$, this reduces to

$$E(M(l)|\mathscr{F}_j) = M(j) = M(\sigma\wedge\tau).$$

If $j > i$, this reduces to

$$E(M(l)|\mathscr{F}_i) = M(i) = M(\sigma\wedge\tau)$$

and since this exhausts all possibilities for values of σ and τ, it follows

$$E(M(\tau)|\mathscr{F}_\sigma) = M(\sigma\wedge\tau) \text{ a.e. } \blacksquare$$

You can also give a version of the above to sub-martingales. This requires the following very interesting decomposition of a sub-martingale into the sum of an increasing stochastic process and a martingale. This was presented earlier as Lemma 29.4.9.

Theorem 31.2.2 *Let $\{X_n\}$ be a sub-martingale. Then there exists a unique stochastic process, $\{A_n\}$ and martingale, $\{M_n\}$ such that*

1. $A_n(\omega) \leq A_{n+1}(\omega)$, $A_1(\omega) = 0$,
2. A_n *is* $\mathscr{F}_{n-1}$ *adapted for all* $n \geq 1$ *where* $\mathscr{F}_0 \equiv \mathscr{F}_1$.

and also $X_n = M_n + A_n$.

Recall that the thing which works is

$$A(n) \equiv \sum_{k=0}^{n-1} E(X(k+1) - X(k) | \mathscr{F}_k),\ A(0) = 0$$

and that this is the only thing which will do what is required.

Now here is a version of the optional sampling theorem for sub-martingales. This was also presented earlier. However, it is good to go through the proof as a review.

Theorem 31.2.3 *Let $\{X(k)\}$ be a real valued sub-martingale with respect to the increasing sequence of σ algebras, $\{\mathscr{F}_k\}$ such that τ is bounded. Then $X(\tau)$ defined as*

$$\omega \to X(\tau(\omega))$$

is integrable and

$$X(\sigma \wedge \tau) \leq E(X(\tau) | \mathscr{F}_\sigma)$$

Proof: That $\omega \to X(\tau(\omega))$ is integrable follows right away as in the optional sampling theorem for martingales. You just consider the finitely many values of τ.

Use Theorem 31.2.2 above to write

$$X(n) = M(n) + A(n)$$

where M is a martingale and A is increasing with $A(n)$ being $\mathscr{F}_{n-1}$ measurable and $A(0) = 0$ as discussed in Theorem 31.2.2. Then

$$E(X(\tau) | \mathscr{F}_\sigma) = E(M(\tau) + A(\tau) | \mathscr{F}_\sigma)$$

Now since A is increasing, you can use the optional sampling theorem for martingales to conclude that, since $\mathscr{F}_{\sigma\wedge\tau} \subseteq \mathscr{F}_\sigma$ and $A(\sigma \wedge \tau)$ is $\mathscr{F}_{\sigma\wedge\tau}$ measurable,

$$\begin{aligned} &\geq E(M(\tau) + A(\sigma \wedge \tau) | \mathscr{F}_\sigma) = E(M(\tau) | \mathscr{F}_\sigma) + A(\sigma \wedge \tau) \\ &= M(\sigma \wedge \tau) + A(\sigma \wedge \tau) = X(\sigma \wedge \tau). \blacksquare \end{aligned}$$

Note that if $\sigma \leq \tau$, the conclusion is $X(\sigma) \leq E(X(\tau) | \mathscr{F}_\sigma)$.

31.3 Doob Optional Sampling Continuous Case

31.3.1 Stopping Times

Let $X(t)$ be a stochastic process adapted to a filtration $\{\mathscr{F}_t\}$ for $t \in [0,T]$ meaning that $X(t)$ is $\mathscr{F}_t$ measurable each $\mathscr{F}_t$ being a subset of $\mathscr{F}$. We will assume two things. The stochastic process is right continuous and the filtration is normal. Recall what this means:

Definition 31.3.1 *A normal filtration is one which satisfies the following :*

1. $\mathscr{F}_0$ *contains all* $A \in \mathscr{F}$ *such that* $P(A) = 0$. *Here* $\mathscr{F}$ *is the* σ *algebra which contains all* $\mathscr{F}_t$.

2. $\mathscr{F}_t = \mathscr{F}_{t+}$ *for all* $t \in I$ *where* $\mathscr{F}_{t+} \equiv \cap_{s>t} \mathscr{F}_s$.

For an $\mathscr{F}$ measurable $[0,\infty)$ valued function τ to be a stopping time, we want to have the stopped process X^τ defined by $X^\tau(t)(\omega) \equiv X(t \wedge \tau(\omega))(\omega)$ to be adapted whenever X is right continuous and adapted. Thus a stopping time is a measurable function which can be used to stop the process while retaining the property of being adapted. The definition of such a condition which will make τ a stopping time is the same as in the case of a discreet process.

Definition 31.3.2 τ *an* $\mathscr{F}$ *measurable function is a stopping time if* $[\tau \le t] \in \mathscr{F}_t$.

Then this definition does what is desired. This is in the following proposition. For convenience, here is a definition.

Definition 31.3.3 *Let* $\{t\}_k \equiv 2^{-k} n$ *where* n *is as large as possible and have* $2^{-k} n \le t$.

It seems like the theory is based on reducing to discrete stopping times defined as follows.

Definition 31.3.4

$$\tau_k(\omega) \equiv \sum_{n=0}^{\infty} \mathscr{X}_{\tau^{-1}\left((n2^{-k},(n+1)2^{-k}]\right)}(\omega)\,(n+1)\,2^{-k}.$$

Thus τ_k *has values in the set* $\left\{n2^{-k}\right\}_{n=0}^{\infty}$, $\tau_k \ge \tau$ *and* τ_k *is within* 2^{-k} *of* τ.

Then τ_k is a discrete stopping time with respect to the increasing σ algebras $\mathscr{F}_{\{t\}_k}$. This is in the following lemma.

Lemma 31.3.5 *Let* τ *be a stopping time and let* τ_k *be defined above in Definition 31.3.4. Then* $[\tau_k \le \{t\}_k] \in \mathscr{F}_{\{t\}_k}$ *and for all* t, $[\tau_k \le t] \in \mathscr{F}_t$. *If you have finitely many stopping times* $\left\{\sigma^k\right\}_{k=1}^n$ *for* $n < \infty$ *then* $\sigma \equiv \min\left\{\sigma^k\right\}_{k=1}^n$ *and* $\sigma \equiv \max\left\{\sigma^k\right\}_{k=1}^n$ *are also stopping times.*

Proof: If $t = (n+1)2^{-k}$ then $[\tau_k \le \{t\}_k]$ is the same as $\left[\tau \le (n+1)2^{-k}\right] = [\tau \le t] \in \mathscr{F}_t = \mathscr{F}_{\{t\}_k}$. The other case is where for some n, $t \in \left(n2^{-k}, (n+1)2^{-k}\right)$. Then $\{t\}_k = n2^{-k}$ and so $[\tau_k \le \{t\}_k]$ is $\left[\tau \le n2^{-k}\right] \in \mathscr{F}_{n2^{-k}} = \mathscr{F}_{\{t\}_k}$. Thus, in particular, $[\tau_k \le t] \in \mathscr{F}_t$ since $\{t\}_k \le t$.

As to the second claim,

$$\left[\min\left\{\sigma^k\right\}_{k=1}^n \le t\right] = \cup_{k=1}^n \left[\sigma^k \le t\right] \in \mathscr{F}_t$$

and $\left[\max\left\{\sigma^k\right\}_{k=1}^n \le t\right] = \cap_{k=1}^n \left[\sigma^k \le t\right] \in \mathscr{F}_t$. ■

Proposition 31.3.6 *Let $\{\mathscr{F}_t\}$ be a normal filtration and let $X(t)$ be a right continuous process adapted to $\{\mathscr{F}_t\}$. Then if τ is a stopping time, it follows that the stopped process X^τ defined by $X^\tau(t) \equiv X(\tau \wedge t)$ is also adapted.*

Proof: Let $\tau_k(\omega) \equiv \sum_{n=0}^{\infty} \mathscr{X}_{\tau^{-1}\left((n2^{-k},(n+1)2^{-k}]\right)}(\omega)(n+1)2^{-k}$. Thus τ_k has discreet values $n2^{-k}, n = 1, \ldots$ and $\tau_k(\omega) = (n+1)2^{-k}$ exactly when ω is in

$$\left[\tau \in \left[0,(n+1)2^{-k}\right]\right] \setminus \left[\tau \in \left[0,n2^{-k}\right]\right]$$

and these sets are both in $\mathscr{F}_{(n+1)2^{-k}}$. Now consider $X(\tau \wedge t)^{-1}(O)$ for O an open set. Since O is open, it follows from right continuity of X that if $X(\tau \wedge t) \in O$, then $X(\tau_k \wedge t) \in O$ whenever k is large enough, depending on ω of course. Thus

$$X(\tau \wedge t)^{-1}(O) = \cup_{m=1}^{\infty} \cap_{k \geq m} X(\tau_k \wedge t)^{-1}(O)$$

Now $X(\tau_k \wedge t)^{-1}(O) = \left([\tau_k \leq \{t\}_k] \cap X(\tau_k)^{-1}(O)\right) \cup \left([\tau_k > \{t\}_k] \cap X(t)^{-1}(O)\right)$. The second term in the union is in $\mathscr{F}_t$ because X is adapted and $[\tau_k > \{t\}_k]$ is the complement of the set $[\tau_k \leq \{t\}_k]$ which is in $\mathscr{F}_t$. The first term is of the form $[\tau_k \leq \{t\}_k] \cap \left(\cup_{j=0}^{2^k\{t\}_k} X\left(2^{-k}j\right)^{-1}(O)\right) \in \mathscr{F}_t$ again because X is adapted. Since each $X(\tau_k \wedge t)^{-1}(O) \in \mathscr{F}_t$, it follows that $X(\tau \wedge t)^{-1}(O) = \cup_{m=1}^{\infty} \cap_{k \geq m} X(\tau_k \wedge t)^{-1}(O)$ is in $\mathscr{F}_t$. ■

By analogy to the discreet case, here are the prior sets.

Definition 31.3.7 *Let $(\Omega, \mathscr{F}, P)$ be a probability space and let $\mathscr{F}_t$ be a filtration. Recall a measurable function, $\tau : \Omega \to [0,\infty]$ is called a stopping time if*

$$[\tau \leq t] \in \mathscr{F}_t$$

for all $t \geq 0$. Associated with a stopping time is the σ algebra, $\mathscr{F}_\tau$ defined by

$$\mathscr{F}_\tau \equiv \{A \in \mathscr{F} : A \cap [\tau \leq t] \in \mathscr{F}_t \text{ for all } t\}.$$

These sets are also called those "prior" to τ.

Note that $\mathscr{F}_\tau$ is obviously closed with respect to countable unions. If $A \in \mathscr{F}_\tau$, then

$$A^C \cap [\tau \leq t] = [\tau \leq t] \setminus (A \cap [\tau \leq t]) \in \mathscr{F}_t$$

Thus $\mathscr{F}_\tau$ is a σ algebra. What if $\tau \leq \sigma$? Does it follow that $\mathscr{F}_\tau \subseteq \mathscr{F}_\sigma$? What about $\omega \to X(\tau)$? Is this measurable?

Recall $\{t\}_k \equiv n2^{-k}$ where n is as large as possible with $n2^{-k} \leq t$.

Proposition 31.3.8 *Let τ be a stopping time and let τ_k be the stopping time having discreet values described in Lemma 31.3.5 which for a given stopping time τ is given by*

$$\tau_k(\omega) \equiv \sum_{n=0}^{\infty} \mathscr{X}_{\tau^{-1}\left((n2^{-k},(n+1)2^{-k}]\right)}(\omega)(n+1)2^{-k}.$$

Then

1. $\mathscr{F}_\tau \subseteq \mathscr{F}_{\tau_k}$ and if $\mathscr{F}_t$ is normal, then if $A \in \mathscr{F}_{\tau_k}$ for all k, it follows that $A \in \mathscr{F}_\tau$.
2. $A \in \mathscr{F}_{\tau_k}$ if and only if $A \cap \left[\tau_k = n2^{-k}\right] \in \mathscr{F}_{n2^{-k}}$ for all n.
3. If $\tau \leq \sigma$, then $\tau_k \leq \sigma_k$.
4. More generally, if $\tau \leq \sigma$ are two stopping times, then $\mathscr{F}_\tau \subseteq \mathscr{F}_\sigma$.
5. If $X(t)$ is a right continuous process adapted to the normal filtration $\mathscr{F}_t$ and τ is a stopping time, then $\omega \to X(\tau(\omega))$ is $\mathscr{F}_\tau$ measurable. Here X has values in a Banach space.

Proof: 1.) Let $A \in \mathscr{F}_\tau$. Then for $t \in (n2^{-k}, (n+1)2^{-k}]$, if $t < (n+1)2^{-k}$,

$$A \cap [\tau_k \leq t] = A \cap \left[\tau_k \leq n2^{-k}\right] = A \cap \left[\tau \leq n2^{-k}\right] \in \mathscr{F}_{n2^{-k}} \subseteq \mathscr{F}_t$$

so $A \in \mathscr{F}_{\tau_k}$. If $t = (n+1)2^{-k}$ then $A \cap [\tau_k \leq t] \in \mathscr{F}_{(n+1)2^{-k}} = \mathscr{F}_t$. Thus $A \in \mathscr{F}_{\tau_k}$ and $\mathscr{F}_\tau \subseteq \mathscr{F}_{\tau_k}$. Now for the other part, Consider $A \cap [\tau \leq t]$. If $t \in (n2^{-k}, (n+1)2^{-k}]$, then $[\tau \leq t] = \left[\tau_k \leq (n+1)2^{-k}\right]$ and so $A \cap [\tau \leq t] = A \cap \left[\tau_k \leq (n+1)2^{-k}\right] \in \mathscr{F}_{(n+1)2^{-k}} \subseteq \mathscr{F}_{t+2^{-k}}$. Since the filtration is normal, and for all k, $A \cap [\tau \leq t] \in \mathscr{F}_{t+2^{-k}}$, it follows that $A \cap [\tau \leq t] \in \mathscr{F}_t$ and so $A \in \mathscr{F}_\tau$ as claimed.

2.) Note

$$\begin{aligned} \left[\tau_k = n2^{-k}\right] &= \left[\tau_k \leq n2^{-k}\right] \setminus \left[\tau_k \leq (n-1)2^{-k}\right] \\ &= \left[\tau \leq 2^{-k}n\right] \setminus \left[\tau \leq (n-1)2^{-k}\right] \end{aligned}$$

If $A \in \mathscr{F}_{\tau_k}$ then by definition,

$$\begin{aligned} A \cap \left[\tau_k = n2^{-k}\right] &= A \cap \left(\left[\tau_k \leq n2^{-k}\right] \setminus \left[\tau_k \leq (n-1)2^{-k}\right]\right) \\ &= A \cap \left(\left[\tau \leq 2^{-k}n\right] \setminus \left[\tau \leq (n-1)2^{-k}\right]\right) \in \mathscr{F}_{n2^{-k}} \end{aligned}$$

Conversely, if

$$A \cap \left[\tau_k = n2^{-k}\right] \in \mathscr{F}_{n2^{-k}}$$

for all n, then consider $A \cap [\tau_k \leq t]$ for $t \in (n2^{-k}, (n+1)2^{-k}]$. Is $A \cap [\tau_k \leq t] \in \mathscr{F}_t$? If $t < (n+1)2^{-k}$, then $A \cap [\tau_k \leq t] = \sum_{j=1}^n A \cap \left[\tau_k = j2^{-k}\right] \in \mathscr{F}_{n2^{-k}} \subseteq \mathscr{F}_t$. If $t = (n+1)2^{-k}$, $A \cap [\tau_k \leq t] = \sum_{j=1}^{n+1} A \cap \left[\tau_k = j2^{-k}\right] \in \mathscr{F}_t$ and so $A \in \mathscr{F}_{(n+1)2^{-k}} \subseteq \mathscr{F}_{\tau_k}$. This proves 2.).

3.) The values of both τ_k and σ_k are $n2^{-k}$ for some nonnegative integer n. τ_k equals $n2^{-k}$ on $\tau^{-1}\left((n-1)2^{-k}, n2^{-k}\right]$. Thus on this set, σ cannot be smaller than or equal to $(n-1)2^{-k}$. Hence σ_k is at least $n2^{-k}$.

4.) Let $\tau \leq \sigma$ and let $A \in \mathscr{F}_\tau$. $A \cap [\sigma \leq t] = A \cap [\tau \leq t] \cap [\sigma \leq t]$ because $\tau \leq \sigma$. However, $A \cap [\tau \leq t] \in \mathscr{F}_t$ and $[\sigma \leq t] \in \mathscr{F}_t$ so the right side is in $\mathscr{F}_t$ which means $A \in \mathscr{F}_\sigma$.

5.)Let U be an open set.

$$X(\tau_k)^{-1}(U) \cap [\tau_k < t] = \cup_{j2^{-k} \leq \{t\}_k} X\left(j2^{-k}\right)^{-1}(U) \in \mathscr{F}_{\{t\}_k}.$$

Now say $t \in \left((n2^{-k}, (n+1)2^{-k}]\right)$. If $t < (n+1)2^{-k}$, then $X(\tau_k)^{-1}(U) \cap [\tau_k \leq t]$ was just given. It is in $\mathscr{F}_{n2^{-k}} = \mathscr{F}_{\{t\}_k} \subseteq \mathscr{F}_t$. Otherwise $t = (n+1)2^{-k}$ and in this case,

$X(\tau_k)^{-1}(U) \cap [\tau_k \leq t]$ is in $\mathscr{F}_{(n+1)2^{-k}} = \mathscr{F}_t$. Thus $\omega \to X(\tau_k)(\omega)$ is $\mathscr{F}_{\tau_k}$ measurable. It follows that for a fixed $\hat{k}$, $X(\tau_k)$ is $\mathscr{F}_{\tau_{\hat{k}}}$ measurable for each $k > \hat{k}$ because τ_k is decreasing in k so this follows from Part 4. Now let $k \to \infty$ and use right continuity of X to conclude that $X(\tau)$ is $\mathscr{F}_{\tau_{\hat{k}}}$ measurable. Thus $X(\tau)^{-1}(U) \in \mathscr{F}_{\tau_{\hat{k}}}$ for each $\hat{k}$ and so, by Part 1, it follows that $X(\tau)^{-1}(U) \in \mathscr{F}_\tau$. Therefore, $X(\tau)$ is $\mathscr{F}_\tau$ measurable. ■

Next is an important proposition which gives a typical example of a stopping time. Since the process has t in an interval, one must be much more careful about the nature of the set which is hit.

Proposition 31.3.9 *Let B be an open subset of topological space E and let $X(t)$ be a right continuous $\mathscr{F}_t$ adapted stochastic process such that $\mathscr{F}_t$ is normal. Then define*

$$\tau(\omega) \equiv \inf\{t > 0 : X(t)(\omega) \in B\}.$$

This is called the first hitting time. Then τ is a stopping time. If $X(t)$ is continuous and adapted to $\mathscr{F}_t$, a normal filtration, then if H is a nonempty closed set such that $H = \cap_{n=1}^{\infty} B_n$ for B_n open, $B_n \supseteq B_{n+1}$,

$$\tau(\omega) \equiv \inf\{t > 0 : X(t)(\omega) \in H\}$$

is also a stopping time.

Proof: Consider the first claim. $\omega \in [\tau = a]$ implies that for each $n \in \mathbb{N}$, there exists $t \in \left[a, a + \frac{1}{n}\right]$ such that $X(t) \in B$. Also for $t < a$, you would need $X(t) \notin B$. By right continuity, this is the same as saying that $X(d) \notin B$ for all rational $d < a$. (If $t < a$, you could let $d_n \downarrow t$ where $X(d_n) \in B^C$, a closed set. Then it follows that $X(t)$ is also in the closed set B^C.) Thus, aside from a set of measure zero, for each $m \in \mathbb{N}$,

$$[\tau = a] = \left(\cap_{n=m}^{\infty} \cup_{t \in \left[a, a+\frac{1}{n}\right]} [X(t) \in B]\right) \cap \left(\cap_{t \in [0,a)} \left[X(t) \in B^C\right]\right)$$

Since $X(t)$ is right continuous, this is the same as

$$\left(\cap_{n=m}^{\infty} \cup_{d \in \mathbb{Q} \cap \left[a, a+\frac{1}{n}\right]} [X(d) \in B]\right) \cap \left(\cap_{d \in \mathbb{Q} \cap [0,a)} \left[X(d) \in B^C\right]\right) \in \mathscr{F}_{a+\frac{1}{m}}$$

Thus, since the filtration is normal,

$$[\tau = a] \in \cap_{m=1}^{\infty} \mathscr{F}_{a+\frac{1}{m}} = \mathscr{F}_{a+} = \mathscr{F}_a$$

I want to consider $[\tau \leq a]$. What of $[\tau < a]$? This is equivalent to saying that $X(t) \in B$ for some $t < a$. Since X is right continuous, this is the same as saying that $X(t) \in B$ for some $t \in \mathbb{Q}, t < a$. Thus

$$[\tau < a] = \cup_{d \in \mathbb{Q}, d < a} [X(d) \in B] \in \mathscr{F}_a$$

It follows that $[\tau \leq a] = [\tau < a] \cup [\tau = a] \in \mathscr{F}_a$. Thus τ is indeed a stopping time.

Now consider the claim involving the additional assumption that $X(t)$ is continuous and it is desired to hit a closed set $H = \cap_{n=1}^{\infty} B_n$ where B_n is open, $B_n \supseteq B_{n+1}$. (Note that if the topological space is a metric space, this is always possible so this is not a big restriction.) Then let τ_n be the first hitting time of B_n by $X(t)$. Then it can be shown that

$$[\tau \leq a] = \cap_n [\tau_n \leq a] \in \mathscr{F}_a$$

To show this, first note that $\omega \in [\tau \leq a]$ if and only if there exists $t \leq a$ such that $X(t)(\omega) \in H$. This follows from continuity and the fact that H is closed. Thus $\tau_n \leq a$ for all n because for some $t \leq a$, $X(t) \in H \subseteq B_n$ for all n. Next suppose $\omega \in [\tau_n \leq a]$ for all n. Then for $\delta_n \downarrow 0$, there exists $t_n \in [0, a+\delta_n]$ such that $X(t_n)(\omega) \in B_n$. It follows there is a subsequence, still denoted by t_n such that $t_n \to t \in [0,a]$. By continuity of X, it must be the case that $X(t)(\omega) \in H$ and so $\omega \in [\tau \leq a]$. This shows the above formula. Now by the first part, each $[\tau_n \leq a] \in \mathscr{F}_a$ and so $[\tau \leq a] \in \mathscr{F}_a$ also. ■

Another useful result for real valued stochastic process is the following in which continuity is generalized to lower semicontinuity.

Proposition 31.3.10 *Let $X(t)$ be a real valued stochastic process which is $\mathscr{F}_t$ adapted for a normal filtration $\mathscr{F}_t$, with the property that $t \to X(t)$ is lower semicontinuous. Then*

$$\tau \equiv \inf\{t : X(t) > \alpha\}$$

is a stopping time.

Proof: As above, for each $m > 0$,

$$[\tau = a] = \left(\cap_{n=m}^{\infty} \cup_{t \in \left[a, a+\frac{1}{n}\right]} [X(t) > \alpha]\right) \cap \left(\cap_{t \in [0,a)} [X(t) \leq \alpha]\right)$$

Now

$$\cap_{t \in [0,a)} [X(t) \leq \alpha] \subseteq \cap_{t \in [0,a), t \in \mathbb{Q}} [X(t) \leq \alpha]$$

If ω is in the right side, then for arbitrary $t < a$, let $t_n \downarrow t$ where $t_n \in \mathbb{Q}$ and $t < a$. Then $X(t) \leq \liminf_{n\to\infty} X(t_n) \leq \alpha$ and so ω is in the left side also. Thus

$$\cap_{t \in [0,a)} [X(t) \leq \alpha] = \cap_{t \in [0,a), t \in \mathbb{Q}} [X(t) \leq \alpha]$$

$$\cup_{t \in \left[a, a+\frac{1}{n}\right]} [X(t) > \alpha] \supseteq \cup_{t \in \left[a, a+\frac{1}{n}\right], t \in \mathbb{Q}} [X(t) > \alpha]$$

If ω is in the left side, then for some t in the given interval, $X(t) > \alpha$. If for all $s \in \left[a, a+\frac{1}{n}\right] \cap \mathbb{Q}$ you have $X(s) \leq \alpha$, then you could take $s_n \to t$ where $X(s_n) \leq \alpha$ and conclude from lower semicontinuity that $X(t) \leq \alpha$ also. Thus there is some rational s where $X(s) > \alpha$ and so the two sides are equal. Hence,

$$[\tau = a] = \left(\cap_{n=m}^{\infty} \cup_{t \in \left[a, a+\frac{1}{n}\right], t \in \mathbb{Q}} [X(t) > \alpha]\right) \cap \left(\cap_{t \in [0,a), t \in \mathbb{Q}} [X(t) \leq \alpha]\right)$$

The first set on the right is in $\mathscr{F}_{a+(1/m)}$ and so is the next set on the right. Hence $[\tau = a] \in \cap_m \mathscr{F}_{a+(1/m)} = \mathscr{F}_a$. To be a stopping time, one needs $[\tau \leq a] \in \mathscr{F}_a$. What of $[\tau < a]$? This equals $\cup_{t \in [0,a)} [X(t) > \alpha] = \cup_{t \in [0,a) \cap \mathbb{Q}} [X(t) > \alpha] \in \mathscr{F}_a$, the equality following from lower semi-continuity. Thus $[\tau \leq a] = [\tau = a] \cup [\tau < a] \in \mathscr{F}_a$. ■

31.3.2 The Optional Sampling Theorem Continuous Case

Proposition 31.3.11 *Let $M(t), t \geq 0$ be a martingale with values in E a separable Banach space and let τ be a bounded stopping time whose maximum is T. Then $M(\tau)$ is $\mathscr{F}_\tau$ measurable and in fact, $\int_\Omega \|M(\tau)\| dP < \infty$. Letting*

$$\tau_k(\omega) \equiv \sum_{n=0}^{\infty} \mathscr{X}_{\tau^{-1}\left((n2^{-k}, (n+1)2^{-k}]\right)}(\omega)(n+1)2^{-k}$$

be the discrete stopping times. $\{M(\tau_k)\}$ are uniformly integrable on $\mathscr{F}_T$.

Proof: From Proposition 31.3.8 $M(\tau)$ is measurable. Since τ is bounded, this is always a finite sum for τ_k. Then each ω is in exactly one $\tau^{-1}\left((n2^{-k},(n+1)2^{-k}]\right)$ for some n. Say $\omega \in \tau^{-1}\left((n2^{-k},(n+1)2^{-k}]\right)$. Then for that $\omega, M(\tau_k(\omega)) = M\left((n+1)2^{-k}\right)(\omega)$. Thus $M(\tau_k(\omega))$ is given by

$$M(\tau_k(\omega)) = \sum_{n=0}^{\infty} \mathcal{X}_{\tau^{-1}\left((n2^{-k},(n+1)2^{-k}]\right)}(\omega) M\left((n+1)2^{-k}\right)(\omega)$$

and

$$\begin{aligned} \|M(\tau_k)(\omega)\| &\leq \sum_{n=0}^{\infty} \mathcal{X}_{\tau^{-1}(I_n)}(\omega) \left\|M\left((n+1)2^{-k}\right)\right\|, \\ I_n &\equiv (n2^{-k},(n+1)2^{-k}] \end{aligned}$$

Then, since $M(t)$ is a martingale, $E\left(M\left((n+1)2^{-k}\right)|\mathcal{F}_{n2^{-k}}\right) = M\left(n2^{-k}\right)$ and so

$$\begin{aligned} \left\|M\left(n2^{-k}\right)\right\| &= \left\|E\left(M\left((n+1)2^{-k}\right)|\mathcal{F}_{n2^{-k}}\right)\right\| \\ &\leq E\left(\left\|M\left((n+1)2^{-k}\right)\right\||\mathcal{F}_{n2^{-k}}\right). \end{aligned}$$

Therefore, iterating this gives

$$\left\|M\left((n+1)2^{-k}\right)\right\| \leq E\left(\|M(T_k)\||\mathcal{F}_{(n+1)2^{-k}}\right)$$

where T_k is the smallest number greater than or equal to T which is of the form $m2^{-k}$ for m a positive integer. Then, since $\mathcal{X}_{\tau^{-1}(I_n)}$ is $\mathcal{F}_{(n+1)2^{-k}}$ measurable, it is $\mathcal{F}_{T_k}$ measurable and so

$$\begin{aligned} \|M(\tau_k)\| &\leq \sum_{n=0}^{\infty} \mathcal{X}_{\tau^{-1}(I_n)}(\omega) E\left(\|M(T_k)\||\mathcal{F}_{(n+1)2^{-k}}\right) \\ &= \sum_{n=0}^{\infty} E\left(\mathcal{X}_{\tau^{-1}(I_n)}\|M(T_k)\||\mathcal{F}_{(n+1)2^{-k}}\right) \end{aligned}$$

because $\mathcal{X}_{\tau^{-1}(I_n)}$ is $\mathcal{F}_{(n+1)2^{-k}}$ measurable. Thus

$$\int_\Omega \|M(\tau_k)\|\,dP \leq \sum_{n=0}^{\infty} \int_\Omega \mathcal{X}_{\tau^{-1}(I_n)}\|M(T_k)\|\,dP = \int_\Omega \|M(T_k)\|\,dP.$$

Thus

$$\int_\Omega \|M(\tau_k)\|\,dP \leq \int_\Omega \|M(T_k)\|\,dP$$

Now use right continuity and Fatou's lemma.

$$\int_\Omega \|M(\tau)\|\,dP \leq \liminf_{k\to\infty} \int_\Omega \|M(\tau_k)\|\,dP \leq \liminf_{k\to\infty} \int_\Omega \|M(T_k)\|\,dP$$

Pick $\hat{T} > T_k$ for all $k = 1,2,\ldots$. Then

$$M(T_k) = E\left(M\left(\hat{T}\right)|\mathcal{F}_{T_k}\right)$$

and so $\|M(T_k)\| \leq E\left(\left\|M\left(\hat{T}\right)\right\| | \mathscr{F}_{T_k}\right)$ and so

$$\int_\Omega \|M(T_k)\| \, dP \leq \int_\Omega E\left(\left\|M\left(\hat{T}\right)\right\| | \mathscr{F}_{T_k}\right) dP = \int_\Omega \left\|M\left(\hat{T}\right)\right\| dP < \infty$$

Therefore,

$$\int_\Omega \|M(\tau)\| \, dP \leq \liminf_{k\to\infty} \int_\Omega \|M(\tau_k)\| \, dP \leq \int_\Omega \left\|M\left(\hat{T}\right)\right\| dP < \infty$$

because it is given that $M(t)$ is in L^1 for each t.

In the above, you could replace Ω with $A \in \mathscr{F}_T$ and conclude

$$\int_A \|M(\tau_k)\| \, dP \leq \int_A \left\|M\left(\hat{T}\right)\right\| dP$$

which implies the $M(\tau_k)$ are uniformly integrable. Given $\varepsilon > 0$ there is $\delta > 0$ such that if $P(A) < \delta$ then

$$\int_A \left\|M\left(\hat{T}\right)\right\| dP < \varepsilon$$

and so also $\int_A \|M(\tau_k)\| \, dP < \varepsilon$. ■

Now consider an increasing in t family of stopping times, $\tau(t)$ $(\omega \to \tau(t)(\omega))$. It turns out this is a sub-martingale.

Lemma 31.3.12 *Let $\{\tau(t)\}$ be an increasing in t family of stopping times, $\tau(t) \geq \tau(s)$ if $s < t$. Then $\tau(t)$ is adapted to the σ algebras $\mathscr{F}_{\tau(t)}$ and $\{\tau(t)\}$ is a sub-martingale adapted to these σ algebras.*

Proof: First I need to show that a stopping time, τ is $\mathscr{F}_\tau$ measurable. Consider $[\tau \leq s]$. Is this in $\mathscr{F}_\tau$? Is $[\tau \leq s] \cap [\tau \leq r] \in \mathscr{F}_r$ for each r? This is obviously so if $s \leq r$ because the intersection reduces to $[\tau \leq s] \in \mathscr{F}_s \subseteq \mathscr{F}_r$. On the other hand, if $s > r$ then the intersection reduces to $[\tau \leq r] \in \mathscr{F}_r$ and so it is clear that τ is $\mathscr{F}_\tau$ measurable. It remains to verify that $t \to \tau(t)$ is a sub-martingale.

Let $s < t$ and let $A \in \mathscr{F}_{\tau(s)}$

$$\int_A E\left(\tau(t) | \mathscr{F}_{\tau(s)}\right) dP \equiv \int_A \tau(t) \, dP \geq \int_A \tau(s) \, dP$$

and this shows $E\left(\tau(t) | \mathscr{F}_{\tau(s)}\right) \geq \tau(s)$ so this is a submartingale as claimed. ■

Now here is an important example. Recall that for τ a stopping time, so is $t \vee \tau$ because

$$[t \vee \tau \leq s] = \begin{matrix}\Omega \text{ if } t \leq s, \emptyset \text{ otherwise} \\ [t \leq s] \end{matrix} \cap [\tau \leq s] \in \mathscr{F}_s.$$

Also recall that if σ is a stopping time, then for adapted Y, $Y(\sigma)$ is $\mathscr{F}_\sigma$ adapted. This is in Proposition 31.3.8.

Proposition 31.3.13 *Let τ be a stopping time and let X be continuous and adapted to the filtration $\mathscr{F}_t$. Then for $a > 0$, define σ as*

$$\sigma(\omega) \equiv \inf\{t > \tau(\omega) : \|X(t)(\omega) - X(\tau(\omega))\| = a\}$$

Then σ is also a stopping time. That is $[\sigma \leq t] \in \mathscr{F}_t$.

Proof: To see this is so, let

$$Y(t)(\omega) = \|X(t \vee \tau)(\omega) - X(\tau(\omega))\|$$

Then $Y(t)$ is $\mathscr{F}_{t\vee\tau}$ measurable. It is desired to show that Y is $\mathscr{F}_t$ adapted. Hence if U is open in $\mathbb{R}$, then

$$Y(t)^{-1}(U) = \left(Y(t)^{-1}(U) \cap [\tau \leq t]\right) \cup \left(Y(t)^{-1}(U) \cap [\tau > t]\right)$$

The second set in the above union on the right equals either $\emptyset$ or $[\tau > t]$ depending on whether $0 \in U$. If $\tau > t$, then $Y(t) = 0$ and so the second set equals $[\tau > t]$ if $0 \in U$. If $0 \notin U$, then the second set equals $\emptyset$. Thus the second set above is in $\mathscr{F}_t$. It is necessary to show the first set is also in $\mathscr{F}_t$. The first set equals

$$Y(t)^{-1}(U) \cap [\tau \leq t] = Y(t)^{-1}(U) \cap [\tau \vee t \leq t]$$

because $[\tau \vee t \leq t] = [\tau \leq t]$. However, $Y(t)^{-1}(U) \in \mathscr{F}_{t\vee\tau}$ and so the set on the right in the above is in $\mathscr{F}_t$. For A to be in $\mathscr{F}_{t\vee\tau}$ means $A \cap [\tau \vee t \leq s] \in \mathscr{F}_s$ for each s. In particular, this is true for $s = t$. Therefore, $Y(t)$ is adapted. Then σ is just the first hitting time for $Y(t)$ to equal the closed set a. Therefore, σ is a stopping time by Proposition 31.3.9. ■

The following corollary involves the same argument. Just replace

$$\|X(t \vee \tau)(\omega) - X(\tau(\omega))\|$$

with $g(X(t \vee \tau)(\omega) - X(\tau(\omega)))$.

Corollary 31.3.14 *Let τ be a stopping time and let X be continuous and adapted to the filtration $\mathscr{F}_t$. Also let g be a continuous real valued function. Then for $a > 0$, define σ as*

$$\sigma(\omega) \equiv \inf\{t > \tau(\omega) : g(X(t)(\omega) - X(\tau(\omega))) = a\}$$

Then σ is also a stopping time.

Next I want a version of the Doob optional sampling theorem which applies to martingales defined on $[0, L], L \leq \infty$. First recall the fundamental property of conditional expectaton that $\|E(f|\mathscr{G})\| \leq E(\|f\||\mathscr{G})$.

Here is a lemma for an optional sampling theorem for the continuous case.

Lemma 31.3.15 *Let $X(t)$ have the property that it is a right continuous nonnegative sub-martingale, $t \geq 0$ such that the filtration $\{\mathscr{F}_t\}$ is normal. Recall this includes $\mathscr{F}_t = \cap_{s>t}\mathscr{F}_s$. Also let τ be a stopping time with values in $[0,T]$. Let $\mathscr{P}_n = \{t_k^n\}_{k=1}^{m_n+1}$ be a sequence of partitions of $[0,T]$ which have the property that*

$$\mathscr{P}_n \subseteq \mathscr{P}_{n+1}, \ \lim_{n\to\infty} \|\mathscr{P}_n\| = 0,$$

where $\|\mathscr{P}_n\| \equiv \sup\left\{\left|t_k^n - t_{k+1}^n\right| : k = 1, 2, \cdots, m_n\right\}$. Then let

$$\tau_n(\omega) \equiv \sum_{k=0}^{m_n} t_{k+1}^n \mathscr{X}_{\tau^{-1}\left((t_k^n, t_{k+1}^n]\right)}(\omega), t_{m_n+1}^n = T$$

It follows that τ_n is a stopping time and also the functions $|X(\tau_n)|$ are uniformly integrable. Furthermore, $|X(\tau)|$ is integrable. Also $X(0), X(\tau_n), X(T)$ is a sub-martingale for the filtration $\mathscr{F}_0, \mathscr{F}_{\tau_n}, \mathscr{F}_T$. If instead $X(t)$ is a martingale having values in a separable Banach space, $X(0), X(\tau_n), X(T)$ is a martingale for the filtration $\mathscr{F}_0, \mathscr{F}_{\tau_n}, \mathscr{F}_T$. In this case, the conclusions about integrability and uniform integrability apply to the sub-martingale $\|X(t)\|$.

Proof: First of all, say $t \in (t_k^n, t_{k+1}^n]$. If $t < t_{k+1}^n$, then

$$[\tau_n \le t] = [\tau \le t_k^n] \in \mathscr{F}_{t_k^n} \subseteq \mathscr{F}_t$$

and if $t = t_{k+1}^n$, then

$$\left[\tau_n \le t_{k+1}^n\right] = \left[\tau \le t_{k+1}^n\right] \in \mathscr{F}_{t_{k+1}^n} = \mathscr{F}_t$$

and so τ_n is a stopping time. Thus from Proposition 31.3.11, $X(\tau_n)$ is in $L^1(\Omega)$ the measurability being resolved from Proposition 31.3.8.

Now

$$X(\tau_n) = X\left(\sum_{k=0}^{m_n} t_{k+1}^n \mathscr{X}_{\tau^{-1}((t_k^n, t_{k+1}^n])}(\omega)\right) = \sum_{k=0}^{m_n} X\left(t_{k+1}^n\right) \mathscr{X}_{\tau^{-1}((t_k^n, t_{k+1}^n])}(\omega)$$

Now $0, \tau_n, T$ is an increasing list of stopping times. Is it the case that it is a sub-martingale for $\mathscr{F}_0, \mathscr{F}_{\tau_n}, \mathscr{F}_T$?

$$E(X(\tau_n)|\mathscr{F}_0) = \sum_{k=0}^{m_n} E\left(X\left(t_{k+1}^n\right) \mathscr{X}_{\tau^{-1}((t_k^n, t_{k+1}^n])}|\mathscr{F}_0\right)$$

$$= \sum_{k=0}^{m_n} \mathscr{X}_{\tau^{-1}((t_k^n, t_{k+1}^n])} E\left(X\left(t_{k+1}^n\right)|\mathscr{F}_0\right) \ge \sum_{k=0}^{m_n} \mathscr{X}_{\tau^{-1}((t_k^n, t_{k+1}^n])} X(0) = X(0)$$

Now also $X(T) = \sum_{k=0}^{m_n} X(T) \mathscr{X}_{\tau^{-1}((t_k^n, t_{k+1}^n])}(\omega)$ so

$$E(X(T)|\mathscr{F}_{\tau_n}) = \sum_{k=0}^{m_n} E\left(\mathscr{X}_{\tau^{-1}((t_k^n, t_{k+1}^n])} X(T)|\mathscr{F}_{\tau_n}\right). \tag{31.1}$$

What is the value of τ_n on $\left[\tau \in (t_k^n, t_{k+1}^n]\right]$? It is t_{k+1}^n, and so from Lemma 31.1.4

$$\begin{aligned} E\left(\mathscr{X}_{\tau^{-1}((t_k^n, t_{k+1}^n])} X(T)|\mathscr{F}_{\tau_n}\right) &= E\left(\mathscr{X}_{\tau^{-1}((t_k^n, t_{k+1}^n])} X(T)|\mathscr{F}_{t_{k+1}^n}\right) \\ &= \mathscr{X}_{\tau^{-1}((t_k^n, t_{k+1}^n])} E\left(X(T)|\mathscr{F}_{t_{k+1}^n}\right) \end{aligned}$$

because $\mathscr{X}_{\tau^{-1}((t_k^n, t_{k+1}^n])}$ is $\mathscr{F}_{t_{k+1}^n}$ measurable. Now since X is a sub-martingale,

$$E\left(X(T)|\mathscr{F}_{t_{k+1}^n}\right) \ge X\left(t_{k+1}^n\right)$$

and so 31.1 implies

$$E(X(T)|\mathscr{F}_{\tau_n}) = \sum_{k=0}^{m_n} E\left(\mathscr{X}_{\tau^{-1}((t_k^n, t_{k+1}^n])} X(T)|\mathscr{F}_{\tau_n}\right)$$

$$\begin{aligned} &= \sum_{k=0}^{m_n} \mathscr{X}_{\tau^{-1}((t_k^n,t_{k+1}^n])} E\left(X(T)\,|\mathscr{F}_{t_{k+1}^n}\right) \\ &\geq \sum_{k=0}^{m_n} \mathscr{X}_{\tau^{-1}((t_k^n,t_{k+1}^n])} X\left(t_{k+1}^n\right) = X(\tau_n) \end{aligned}$$

Thus this is indeed a sub-martingale. The same argument holds in case X is a martingale. One simply replaces the inequalities with equal signs.

Having shown that $X(0), X(\tau_n), X(T)$ is a sub-martingale,

$$\begin{aligned} \int_{[X(\tau_n)\geq\lambda]} X(\tau_n)\,dP &\leq \int_{[X(\tau_n)\geq\lambda]} E(X(T)\,|\mathscr{F}_{\tau_n})\,dP \\ &= \int_\Omega E\left(\mathscr{X}_{[X(\tau_n)\geq\lambda]} X(T)\,|\mathscr{F}_{\tau_n}\right) dP \\ &= \int_{[X(\tau_n)\geq\lambda]} X(T)\,dP \end{aligned}$$

If the interest were in a martingale where $X(t)$ is in a Banach space, you would simply do all the remaining analysis for the sub-martingale $\|X(t)\|$. Thus, from now on, I will mainly consider a real sub-martingale. From maximal estimates, for example Theorem 29.3.14,

$$P([X(\tau_n)\geq\lambda]) \leq \frac{1}{\lambda}\int_\Omega X(T)^+\,dP = \frac{1}{\lambda}\int_\Omega X(T)\,dP$$

and now it follows from the above that the random variables $X(\tau_n)$ are equiintegrable. Recall this means that

$$\lim_{\lambda\to\infty}\sup_n \int_{[X(\tau_n)\geq\lambda]} X(\tau_n)\,dP = 0$$

Hence they are uniformly integrable and bounded in L^1.

To verify again that $|X(\tau)|$ is integrable, note that by right continuity, $X(\tau_n)\to X(\tau)$ pointwise. Apply the Vitali convergence theorem to obtain

$$\int_\Omega |X(\tau)|\,dP = \lim_{n\to\infty}\int_\Omega |X(\tau_n)|\,dP \leq \int_\Omega X(T)\,dP < \infty. \blacksquare$$

Theorem 31.3.16 *Let $M(t)$ be a right continuous martingale with values in a separable Banach space adapted to a normal filtration. Let σ,τ be two stopping times such that τ is bounded. Then $M(\sigma\wedge\tau) = E(M(\tau)\,|\mathscr{F}_\sigma)$. If X is a real sub-martingale, $X(\sigma\wedge\tau)\leq E(X(\tau)\,|\mathscr{F}_\sigma)$.*

Proof: Letting $M(t), t\geq 0$ be a martingale with values in a separable Banach space adapted to a filtration $\mathscr{F}_t$. Let τ_k and σ_k be the discreet stopping times such that τ_k is at least as big as τ but within 2^{-k} of τ discussed earlier. Therefore, from the optional sampling theorem for discreet martingales in Theorem 31.2.1,

$$M(\sigma_n\wedge\tau_n) = E(M(\tau_n)\,|\mathscr{F}_{\sigma_n})$$

Now let $A\in\mathscr{F}_\sigma$. Using Proposition 31.3.8 as needed, $\mathscr{F}_\sigma\subseteq\mathscr{F}_{\sigma_n}$ and

$$\int_A M(\sigma_n\wedge\tau_n)\,dP = \int_A E(M(\tau_n)\,|\mathscr{F}_{\sigma_n})\,dP \overset{\mathscr{F}_\sigma\subseteq\mathscr{F}_{\sigma_n}}{=} \int_A M(\tau_n)\,dP$$

Next note that from the right continuity of M, $M(\sigma_n \wedge \tau_n) \to M(\sigma \wedge \tau)$ and $M(\tau_n) \to M(\tau)$ and so, by uniform integrability from Proposition $\mathcal{AMS}$-LaTeX, the Vitali convergence theorem applies and we conclude that on passing to a limit,

$$\int_A M(\sigma \wedge \tau)\, dP = \int_A M(\tau)\, dP$$

Since $M(\sigma \wedge \tau)$ is $\mathscr{F}_{\sigma\wedge\tau}$ measurable, hence $\mathscr{F}_\sigma$ measurable, and so, from the definition of conditional expectation, the fact that A is arbitrary implies $M(\sigma \wedge \tau) = E(M(\tau)\,|\mathscr{F}_\sigma)$.

Now consider the case where X is a sub-martingale. Then by the same observation above about these stopping times and the discreet theory,

$$E(X(\tau_n)\,|\mathscr{F}_{\sigma_n}) \geq X(\tau_n \wedge \sigma_n)$$

and so, if $A \in \mathscr{F}_\sigma$

$$\int_A X(\tau_n \wedge \sigma_n)\, dP \leq \int_A E(X(\tau_n)\,|\mathscr{F}_{\sigma_n})\, dP \equiv \int_A X(\tau_n)\, dP$$

and so, by right continuity and Lemma 31.3.15 and the Vitali convergence theorem, we can pass to a limit and conclude that

$$\int_A X(\tau \wedge \sigma)\, dP \leq \int_A X(\tau)\, dP = \int_A E(X(\tau)\,|\mathscr{F}_\sigma)\, dP$$

Now $X(\tau \wedge \sigma)$ is $\mathscr{F}_{\tau\wedge\sigma}$ measurable so this function is also $\mathscr{F}_\sigma$ measurable and so, since the above inequality holds for all $A \in \mathscr{F}_\sigma$, it follows that $X(\tau \wedge \sigma) \leq E(X(\tau)\,|\mathscr{F}_\sigma)$. ■

Note that a function defined on a countable ordered set such as the integers or equally spaced points is right continuous so the optional sampling theorem for discreet processes is a special case of this one.

31.4 Maximal Inequalities and Stopping Times

As in the case of discrete martingales and sub-martingales, there are maximal inequalities available. Typical ones were presented earlier but here I will use the idea of a stopping time. This gives a typical applicaton of stopping times by making it possible to consider a bounded process and do all the hard work with it and then pass to a limit.

Definition 31.4.1 *Let $X(t)$ be a right continuous sub-martingale for $t \in I$ and let $\{\tau_n\}$ be a sequence of stopping times such that $\lim_{n\to\infty} \tau_n = \infty$. Then X^{τ_n} is called the stopped sub-martingale and it is defined by*

$$X^{\tau_n}(t) \equiv X(t \wedge \tau_n).$$

More generally, if τ is a stopping time, the stopped sub-martingale (martingale) is $X^\tau(t) \equiv X(t \wedge \tau)$.

Proposition 31.4.2 *The stopped sub-martingale is a sub-martingale.*

Proof: By the optional sampling theorem for sub-martingales, Theorem 31.3.16, it follows that for $s < t$,

$$\begin{aligned} E(X^\tau(t)\,|\mathscr{F}_s) &\equiv E(X(t \wedge \tau)\,|\mathscr{F}_s) \geq X(t \wedge \tau \wedge s) \\ &= X(\tau \wedge s) \equiv X^\tau(s). \blacksquare \end{aligned}$$

Note that a similar argument would work for martingales.

Theorem 31.4.3 *Let $\{X(t)\}$ be a right continuous nonnegative sub-martingale adapted to the normal filtration $\mathscr{F}_t$ for $t \in [0,T]$. Let $p \geq 1$. Define*

$$X^*(t) \equiv \sup\{X(s) : 0 < s < t\},\ X^*(0) \equiv 0.$$

Then for $\lambda > 0$, if $X(t)^p$ is in $L^1(\Omega)$ for each t,

$$P([X^*(T) > \lambda]) \leq \frac{1}{\lambda^p} \int \mathscr{X}_{[X^*(T)>\lambda]} X(T)^p \, dP \tag{31.2}$$

If $X(t)$ is continuous, the above inequality holds without this assumption. In case $p > 1$, and $X(t)$ continuous, then for each $t \leq T$,

$$\left(\int_\Omega |X^*(t)|^p \, dP \right)^{1/p} \leq \frac{p}{p-1} \left(\int_\Omega X(T)^p \, dP \right)^{1/p} \tag{31.3}$$

Proof: The first inequality follows from Theorem 30.5.2. However, it can also be obtained a different way using stopping times. First note that from right continuity, $X^*(t) = \sup\{X(d) : d \in D\}$ where D is a dense countable set in $(0,t)$. Therefore, $X^*(t)$ is always measurable.

First I will assume $X(t)$ is a bounded sub-martingale. These certainly exist. Just take a bounded stopping time τ and consider X^τ.

Define the stopping time

$$\tau \equiv \inf\{t > 0 : X(t) > \lambda\} \wedge T.$$

(The infimum over an empty set will equal ∞.)This is a stopping time by 31.3.9 because it is just a continuous function of the first hitting time of an open set. Also from the definition of X^* in which the supremum is taken over an open interval,

$$[\tau < t] = [X^*(t) > \lambda]$$

Note this also shows $X^*(t)$ is $\mathscr{F}_t$ measurable. Then it follows that $X^p(t)$ is also a sub-martingale since r^p is increasing and convex. By the optional sampling theorem,

$$X(0)^p, X(\tau)^p, X(T)^p$$

is a sub-martingale. Recall $X(\sigma \wedge \tau) \leq E(X(\tau)|\mathscr{F}_\sigma)$ when τ is bounded. I need to verify that

$$E(X(T)^p|\mathscr{F}_\tau) \geq X(\tau)^p, E(X(\tau)^p|\mathscr{F}_0) \geq X(0)^p.$$

But from the optional sampling theorem Theorem 31.3.16

$$\begin{array}{rcl} E(X(T)^p|\mathscr{F}_\tau) & \geq & X(T \wedge \tau)^p = X(\tau)^p \\ E(X(\tau)^p|\mathscr{F}_0) & \geq & X(\tau \wedge 0)^p = X(0)^p \end{array}$$

Also $[\tau < T] \in \mathscr{F}_\tau$. Recall that A is $\mathscr{F}_\tau$ measurable means $A \cap [\tau \leq t] \in \mathscr{F}_t$. Since τ is a stopping time, $[\tau \leq T] \cap [\tau \leq t] = [\tau \leq t] \in \mathscr{F}_t$ and so $[\tau \leq T] \in \mathscr{F}_\tau$.

$$\int_{[\tau<T]} X(\tau)^p \, dP \leq \int_{[\tau<T]} E(X(T)^p|\mathscr{F}_\tau) \, dP = \int_{[\tau<T]} X(T)^p \, dP$$

By right continuity, on $[\tau < T]$, $X(\tau) \geq \lambda$. Therefore,

$$\begin{aligned} \lambda^p P([X^*(T) > \lambda]) &= \lambda^p P([\tau < T]) \\ \overset{X(\tau)\geq\lambda}{\leq} \int_{[\tau<T]} X(\tau)^p \, dP &= \int_{[\tau<T]} E(X(T)^p \,|\mathscr{F}_\tau)\, dP = \int_{[X^*(T)>\lambda]} X(T)^p \, dP \end{aligned}$$

This proves 31.2 in case X is bounded. In general case, suppose X is not just right continuous but also continuous.

Next let $\{\tau_n\}$ be a "localizing sequence" given by

$$\tau_n \equiv \inf\{t : X(t) > n\}.$$

If $t < \tau_n$, then $X(t) \leq n$ by definition of τ_n. Could $X(\tau_n) > n$? If so, then by continuity, $X(t) > n$ for some $t < \tau_n$ so τ_n was not chosen correctly. Thus X^{τ_n} is bounded because $X(\tau_n \wedge t) \leq n$, and so from what was just shown,

$$\lambda^p P\left(\left[(X^{\tau_n})^*(T) > \lambda\right]\right) \leq \int_{\left[(X^{\tau_n})^*(T)>\lambda\right]} (X^{\tau_n})(T)^p \, dP$$

Then $(X^{\tau_n})(T)$ is increasing to $X(T)$ and $(X^{\tau_n})^*(T)$ increases to $X^*(T)$ as $n \to \infty$ so 31.2 follows from the monotone convergence theorem. This proves 31.2.

Let X^{τ_n} be as just defined. Thus it is a bounded sub-martingale. To save on notation, the X in the following argument is really X^{τ_n}. This is done so that all the integrals are finite. If $p > 1$, then from the first part,

$$\int_\Omega |X^*(t)|^p \, dP \leq \int_\Omega |X^*(T)|^p \, dP = \int_0^\infty p\lambda^{p-1} \overbrace{P([X^*(T) > \lambda])}^{\leq \frac{1}{\lambda}\int \mathscr{X}_{[X^*(T)>\lambda]} X(T) dP} \, d\lambda$$

$$\leq p \int_0^\infty \lambda^{p-1} \frac{1}{\lambda} \int_\Omega \mathscr{X}_{[X^*(T)>\lambda]} X(T) \, dP d\lambda$$

By Lemma 29.3.13, applied to the second half of the above and using Holder's inequality,

$$\begin{aligned} \int_\Omega |X^*(T)|^p \, dP &\leq p \int_\Omega X(T) \int_0^{X^*(T)} \lambda^{p-2} d\lambda \, dP = p \int_\Omega X(T) \frac{X^*(T)^{p-1}}{p-1} dP \\ &\leq \frac{p}{p-1} \left(\int_\Omega X^*(T)^p \, dP \right)^{1/p'} \left(\int_\Omega X(T)^p \, dP \right)^{1/p} \end{aligned}$$

Now divide both sides by $(\int_\Omega X^*(T)^p \, dP)^{1/p'}$ and restore X^{τ_n} for X.

$$\left(\int_\Omega X^{\tau_n *}(T)^p \, dP \right)^{1/p} \leq \frac{p}{p-1} \left(\int_\Omega X^{\tau_n}(T)^p \, dP \right)^{1/p}$$

Now let $n \to \infty$ and use the monotone convergence theorem to obtain the inequality of the theorem 31.3. ■

Here is another sort of maximal inequality in which $X(t)$ is not assumed nonnegative. A version of this was also presented earlier.

Theorem 31.4.4 *Let $\{X(t)\}$ be a right continuous sub-martingale adapted to the normal filtration $\mathcal{F}_t$ for $t \in [0,T]$ and $X^*(t)$ defined as in Theorem 31.4.3*

$$X^*(t) \equiv \sup\{X(s) : 0 < s < t\},\ X^*(0) \equiv 0,$$

$$P([X^*(T) > \lambda]) \leq \frac{1}{\lambda} E(|X(T)|) \tag{31.4}$$

For $t > 0$, let

$$X_*(t) = \inf\{X(s) : s < t\}.$$

Then

$$P([X_*(T) < -\lambda]) \leq \frac{1}{\lambda} E(|X(T)| + |X(0)|) \tag{31.5}$$

Also

$$P([\sup\{|X(s)| : s < T\} > \lambda])$$
$$\leq \frac{2}{\lambda} E(|X(T)| + |X(0)|) \tag{31.6}$$

Proof: The function $f(r) = r^+ \equiv \frac{1}{2}(|r| + r)$ is convex and increasing. Therefore, $X^+(t)$ is also a sub-martingale but this one is nonnegative. Also

$$[X^*(T) > \lambda] = \left[\left(X^+\right)^*(T) > \lambda\right]$$

and so from Theorem 31.4.3,

$$P([X^*(T) > \lambda]) = P\left(\left[\left(X^+\right)^*(T) > \lambda\right]\right) \leq \frac{1}{\lambda} E\left(X^+(T)\right) \leq \frac{1}{\lambda} E(|X(T)|).$$

Next let

$$\tau = \min(\inf\{t : X(t) < -\lambda\}, T)$$

then as before, $X(0), X(\tau), X(T)$ is a sub-martingale and so

$$\int_{[\tau<T]} X(\tau)\, dP + \int_{[\tau=T]} X(\tau)\, dP = \int_\Omega X(\tau)\, dP \geq \int_\Omega X(0)\, dP$$

Now for $\omega \in [\tau < T], X(t)(\omega) < -\lambda$ for some $t < T$ and so by right continuity, $X(\tau)(\omega) \leq -\lambda$. therefore,

$$-\lambda \int_{[\tau<T]} dP \geq -\int_{[\tau=T]} X(T)\, dP + \int_\Omega X(0)\, dP$$

If $X_*(T) < -\lambda$, then from the definition given above, there exists $t < T$ such that $X(t) < -\lambda$ and so $\tau < T$. If $\tau < T$, then by definition, there exists $t < T$ such that $X(t) < -\lambda$ and so $X_*(T) < -\lambda$. Hence $[\tau < T] = [X_*(T) < -\lambda]$. It follows that

$$P([X_*(T) < -\lambda]) = P([\tau < T])$$

$$\leq \frac{1}{\lambda} \int_{[\tau=T]} X(T)\, dP - \frac{1}{\lambda} \int_\Omega X(0)\, dP \leq \frac{1}{\lambda} E(|X(T)| + |X(0)|)$$

and this proves 31.5.

Finally, combining the above two inequalities,

$$P([\sup\{|X(s)| : s < T\} > \lambda]) = P([X_*(T) < -\lambda]) + P([X^*(T) > \lambda])$$

$$\leq \frac{2}{\lambda} E(|X(T)| + |X(0)|). \blacksquare$$

31.5 Continuous Sub-martingale Convergence

Here, $\{Y(t)\}$ will be a continuous sub-martingale and $a<b$. Let $X(t)\equiv (Y(t)-a)_+ + a$ so $X(0)\geq a$. Then X is also a sub-martingale. It is an increasing convex function of one. If $Y(t)$ has an upcrossing of $[a,b]$, then $X(t)$ starts off at a and ends up at least as large as b. If $X(t)$ has an upcrossing of $[a,b]$ then it must start off at a since it cannot be smaller and it ends up at least as large as b. Thus we can count the upcrossings of $Y(t)$ by considering the upcrossings of $X(t)$ and $X(t)$ is always at least as large as a.

The next task is to consider an upcrossing estimate as was done before for discrete sub-martingales.

$$\begin{aligned}
\tau_0 &\equiv \min\left(\inf\{t>0 : X(t)=a\}, M\right),\\
\tau_1 &\equiv \min\left(\inf\left\{t>0 : (X(t\vee\tau_0)-X(\tau_0))_+ = b-a\right\}, M\right),\\
\tau_2 &\equiv \min\left(\inf\left\{t>0 : (X(\tau_1)-X(t\vee\tau_1))_+ = b-a\right\}, M\right),\\
\tau_3 &\equiv \min\left(\inf\left\{t>0 : (X(t\vee\tau_2)-X(\tau_2))_+ = b-a\right\}, M\right),\\
\tau_4 &\equiv \min\left(\inf\left\{t>0 : (X(\tau_3)-X(t\vee\tau_3))_+ = b-a\right\}, M\right),\\
&\vdots
\end{aligned}$$

If $X(t)$ is never a, then $\tau_0\equiv M$ where we assume $t\in[0,M]$ and there are no upcrossings. It is obvious $\tau_1\geq\tau_0$ since otherwise, the inequality could not hold. Thus the evens have $X(\tau_{2k})=a$ and $X(\tau_{2k+1})=b$. The following lemma follows from Corollary 31.3.14.

Lemma 31.5.1 *The above τ_i are stopping times for $t\in[0,M]$.*

Note that in the above, if $\eta=M$, then $\sigma=M$ also. Thus in the definition of the τ_i, if any $\tau_i=M$, it follows that also $\tau_{i+1}=M$ and so there is no change in the stopping times. Also note that these stopping times τ_i are increasing as i increases.

Let

$$U^{nM}_{[a,b]}\equiv\lim_{\varepsilon\to 0}\sum_{k=0}^{n}\frac{X(\tau_{2k+1})-X(\tau_{2k})}{\varepsilon+X(\tau_{2k+1})-X(\tau_{2k})}$$

Note that if an upcrossing occurs after τ_{2k} on $[0,M]$, then $\tau_{2k+1}>\tau_{2k}$ because there exists t such that

$$(X(t\vee\tau_{2k})-X(\tau_{2k}))_+ = b-a$$

However, you could have $\tau_{2k+1}>\tau_{2k}$ without an upcrossing occuring. This happens when $\tau_{2k}<M$ and $\tau_{2k+1}=M$ which may mean that $X(t)$ never again climbs to b. You break the sum into those terms where $X(\tau_{2k+1})-X(\tau_{2k})=b-a$ and those where this is less than $b-a$. Suppose for a fixed ω, the terms where the difference is $b-a$ are for $k\leq m$. Then there might be a last term for which $X(\tau_{2k+1})-X(\tau_{2k})<b-a$ because it fails to complete the up crossing. There is only one of these at $k=m+1$. Then the above sum is

$$\begin{aligned}
&\leq \frac{1}{b-a}\sum_{k=0}^{m}X(\tau_{2k+1})-X(\tau_{2k})+\frac{X(M)-a}{\varepsilon+X(M)-a}\\
&\leq \frac{1}{b-a}\sum_{k=0}^{n}X(\tau_{2k+1})-X(\tau_{2k})+\frac{X(M)-a}{\varepsilon+X(M)-a}\\
&\leq \frac{1}{b-a}\sum_{k=0}^{n}X(\tau_{2k+1})-X(\tau_{2k})+1
\end{aligned}$$

Then $U_{[a,b]}^{nM}$ is clearly a random variable which is at least as large as the number of upcrossings occurring for $t \leq M$ using only $2n+1$ of the stopping times. From the optional sampling theorem,

$$\begin{aligned} E(X(\tau_{2k})) - E(X(\tau_{2k-1})) &= \int_\Omega X(\tau_{2k}) - X(\tau_{2k-1})\, dP \\ &= \int_\Omega E\left(X(\tau_{2k}) | \mathscr{F}_{\tau_{2k-1}}\right) - X(\tau_{2k-1})\, dP \\ &\geq \int_\Omega X(\tau_{2k-1}) - X(\tau_{2k-1})\, dP = 0 \end{aligned}$$

Note that $X(\tau_{2k}) = a$ while $X(\tau_{2k-1}) = b$ so the above may seem surprising. However, the two stopping times can both equal M so this is actually possible. For example, it could happen that $X(t) = a$ for all $t \in [0,M]$.

Next, take the expectation of both sides,

$$E\left(U_{[a,b]}^{nM}\right) \leq \frac{1}{b-a} \sum_{k=0}^{n} E(X(\tau_{2k+1})) - E(X(\tau_{2k})) + 1$$

$$\leq \frac{1}{b-a} \sum_{k=0}^{n} E(X(\tau_{2k+1})) - E(X(\tau_{2k})) + \frac{1}{b-a} \sum_{k=1}^{n} E(X(\tau_{2k})) - E(X(\tau_{2k-1})) + 1$$

$$= \frac{1}{b-a} (E(X(\tau_1)) - E(X(\tau_0))) + \frac{1}{b-a} \sum_{k=1}^{n} E(X(\tau_{2k+1})) - E(X(\tau_{2k-1})) + 1$$

$$\begin{aligned} &\leq \frac{1}{b-a} (E(X(\tau_{2n+1})) - E(X(\tau_0))) + 1 \\ &\leq \frac{1}{b-a} (E(X(M)) - a) + 1 \end{aligned}$$

which does not depend on n. The last inequality follows because $0 \leq \tau_{2n+1} \leq M$ and $X(t)$ is a sub-martingale. Let $n \to \infty$ to obtain

$$E\left(U_{[a,b]}^{M}\right) \leq \frac{1}{b-a} (E(X(M)) - a) + 1$$

where $U_{[a,b]}^{M}$ is an upper bound to the number of upcrossings of $\{X(t)\}$ on $[0,M]$. This proves the following interesting upcrossing estimate.

Lemma 31.5.2 *Let $\{Y(t)\}$ be a continuous sub-martingale adapted to a normal filtration $\mathscr{F}_t$ for $t \in [0,M]$. Then if $U_{[a,b]}^{M}$ is defined as the above upper bound to the number of upcrossings of $\{Y(t)\}$ for $t \in [0,M]$, then this is a random variable and*

$$\begin{aligned} E\left(U_{[a,b]}^{M}\right) &\leq \frac{1}{b-a} \left(E(Y(M) - a)_+ + a - a\right) + 1 \\ &= \frac{1}{b-a} E(|Y(M)|) + \frac{1}{b-a} |a| + 1 \end{aligned}$$

With this it is easy to prove a continuous sub-martingale convergence theorem.

Theorem 31.5.3 *Let $\{X(t)\}$ be a continuous sub-martingale adapted to a normal filtration such that*

$$\sup_t \{E(|X(t)|)\} = C < \infty.$$

Then there exists $X_\infty \in L^1(\Omega)$ such that

$$\lim_{t\to\infty} X(t)(\omega) = X_\infty(\omega) \text{ a.e. } \omega.$$

Proof: Let $U_{[a,b]}$ be defined by

$$U_{[a,b]} = \lim_{M\to\infty} U^M_{[a,b]}.$$

Thus the random variable $U_{[a,b]}$ is an upper bound for the number of upcrossings. From Lemma 31.5.2 and the assumption of this theorem, there exists a constant C independent of M such that

$$E\left(U^M_{[a,b]}\right) \le \frac{C}{b-a} + 1.$$

Letting $M \to \infty$, it follows from monotone convergence theorem that

$$E\left(U_{[a,b]}\right) \le \frac{C}{b-a} + 1$$

also. Therefore, there exists a set of measure 0 N_{ab} such that if $\omega \notin N_{ab}$, then $U_{[a,b]}(\omega) < \infty$. That is, there are only finitely many upcrossings. Now let

$$N = \cup\{N_{ab} : a,b \in \mathbb{Q}\}.$$

It follows that for $\omega \notin N$, it cannot happen that

$$\limsup_{t\to\infty} X(t)(\omega) - \liminf_{t\to\infty} X(t)(\omega) > 0$$

because if this expression is positive, there would be arbitrarily large values of t where $X(t)(\omega) > b$ and arbitrarily large values of t where $X(t)(\omega) < a$ where a,b are rational numbers chosen such that

$$\limsup_{t\to\infty} X(t)(\omega) > b > a > \liminf_{t\to\infty} X(t)(\omega)$$

Thus there would be infinitely many upcrossings which is not allowed for $\omega \notin N$. Therefore, the limit $\lim_{t\to\infty} X(t)(\omega)$ exists for a.e. ω. Let $X_\infty(\omega)$ equal this limit for $\omega \notin N$ and let $X_\infty(\omega) = 0$ for $\omega \in N$. Then X_∞ is measurable and by Fatou's lemma,

$$\int_\Omega |X_\infty(\omega)|\, dP \le \liminf_{n\to\infty} \int_\Omega |X(n)(\omega)|\, dP < C. \blacksquare$$

Now here is an interesting result of Doob.

Theorem 31.5.4 *Let $\{M(t)\}$ be a continuous real martingale adapted to the normal filtration $\mathcal{F}_t$. Then the following are equivalent.*

1. *The random variables $M(t)$ are equi-integrable.*

2. *There exists* $M(\infty) \in L^1(\Omega)$ *such that* $\lim_{t\to\infty} \|M(\infty) - M(t)\|_{L^1(\Omega)} = 0$.

In this case, $M(t) = E(M(\infty)|\mathscr{F}_t)$ *and convergence also takes place pointwise.*

Proof: Suppose the equi-integrable condition. Then there exists λ large enough that for all t,

$$\int_{[|M(t)|\geq\lambda]} |M(t)|\,dt < 1.$$

It follows that for all t,

$$\begin{aligned}\int_\Omega |M(t)|\,dP &= \int_{[|M(t)|\geq\lambda]} |M(t)|\,dP + \int_{[|M(t)|<\lambda]} |M(t)|\,dP \\ &\leq 1+\lambda.\end{aligned}$$

Since the martingale is bounded in L^1, by Theorem 31.5.3 there exists $M(\infty) \in L^1(\Omega)$ such that $\lim_{t\to\infty} M(t)(\omega) = M(\infty)(\omega)$ pointwise a.e. By the assumption $\{M(t)\}$ are equi-integrable, it follows from Proposition 10.9.6 these functions are uniformly integrable. Then by the Vitali convergence theorem, Theorem 10.9.7, if $t_n \to \infty$, then

$$\|M(t_n) - M(\infty)\|_{L^1(\Omega)} \to 0$$

Next suppose there is a function $M(\infty)$ to which $M(t)$ converges in $L^1(\Omega)$. Then for t fixed and $A \in \mathscr{F}_t$, then as $s \to \infty, s > t$

$$\begin{aligned}\int_A M(t)\,dP &= \int_A E(M(s)|\mathscr{F}_t)\,dP \equiv \int_A M(s)\,dP \\ &\to \int_A M(\infty)\,dP = \int_A E(M(\infty)|\mathscr{F}_t)\end{aligned}$$

which shows $E(M(\infty)|\mathscr{F}_t) = M(t)$ a.e. since $A \in \mathscr{F}_t$ is arbitrary. By Theorem 24.12.1,

$$\begin{aligned}\int_{[|M(t)|\geq\lambda]} |M(t)|\,dP &= \int_{[|M(t)|\geq\lambda]} |E(M(\infty)|\mathscr{F}_t)|\,dP \\ &\leq \int_{[|M(t)|\geq\lambda]} E(|M(\infty)||\mathscr{F}_t)\,dP \\ &= \int_{[|M(t)|\geq\lambda]} |M(\infty)|\,dP \qquad (31.7)\end{aligned}$$

Now from this,

$$\begin{aligned}\lambda P([|M(t)|\geq\lambda]) &\leq \int_{[|M(t)|\geq\lambda]} |M(t)|\,dP \leq \int_\Omega |E(M(\infty)|\mathscr{F}_t)|\,dP \\ &\leq \int_\Omega E(|M(\infty)||\mathscr{F}_t)\,dP = \int_\Omega |M(\infty)|\,dP\end{aligned}$$

and so

$$P([|M(t)|\geq\lambda]) \leq \frac{C}{\lambda}$$

From 31.7, this shows $\{M(t)\}$ is equi-integrable hence uniformly integrable because this is true of the single function $|M(\infty)|$. ■

31.6 Hitting This Before That

Let $\{M(t)\}$ be a real valued continuous martingale for $t \in [0,T]$ where $T \leq \infty$ and $M(0) = 0$. In case $T = \infty$, assume the conditions of Theorem 31.5.4 are satisfied. Thus, according to these conditions, there exists $M(\infty)$ and the $M(t)$ are equi-integrable. With the Doob optional sampling theorem it is possible to estimate the probability that $M(t)$ hits a before it hits b where $a < 0 < b$. There is no loss of generality in assuming $T = \infty$ since if it is less than ∞, you could just let $M(t) \equiv M(T)$ for all $t > T$. In this case, the equiintegrability of the $M(t)$ follows because for $t < T$,

$$\int_{[|M(t)|>\lambda]} |M(t)|\,dP = \int_{[|M(t)|>\lambda]} |E(M(T)|\mathscr{F}_t)|\,dP$$

$$\leq \int_{[|M(t)|>\lambda]} E(|M(T)||\mathscr{F}_t)\,dP = \int_{[|M(t)|>\lambda]} |M(T)|\,dP$$

and from Theorem 31.4.4,

$$P(|M(t)| > \lambda) \leq P([M^*(t) > \lambda]) \leq \frac{1}{\lambda}\int_\Omega |M(T)|\,dP.$$

Definition 31.6.1 *Let M be a process adapted to the filtration $\mathscr{F}_t$ and let τ be a stopping time. Then M^τ, called the stopped process is defined by*

$$M^\tau(t) \equiv M(\tau \wedge t).$$

With this definition, here is a simple lemma. I will use this lemma whenever convenient without comment.

Lemma 31.6.2 *Let M be a right continuous martingale adapted to the normal filtration $\mathscr{F}_t$ and let τ be a stopping time. Then M^τ is also a martingale adapted to the filtration $\mathscr{F}_t$. The same is true for a sub-martingale.*

Proof: Let $s < t$. By the Doob optional sampling theorem,

$$E(M^\tau(t)|\mathscr{F}_s) \equiv E(M(\tau \wedge t)|\mathscr{F}_s) = M(\tau \wedge t \wedge s) = M^\tau(s).$$

As for a sub-martingale $X(t)$, for $s < t$

$$E(X^\tau(t)|\mathscr{F}_s) \equiv E(X(\tau \wedge t)|\mathscr{F}_s) \geq X(\tau \wedge t \wedge s) \equiv X^\tau(s). \blacksquare$$

Theorem 31.6.3 *Let $\{M(t)\}$ be a continuous real valued martingale adapted to the normal filtration $\mathscr{F}_t$ and let*

$$M^* \equiv \sup\{|M(t)| : t \geq 0\}$$

and $M(0) = 0$. Letting

$$\tau_x \equiv \inf\{t > 0 : M(t) = x\}$$

Then if $a < 0 < b$ the following inequalities hold.

$$(b-a)P([\tau_b \leq \tau_a]) \geq -aP([M^* > 0]) \geq (b-a)P([\tau_b < \tau_a])$$

and

$$(b-a)P([\tau_a < \tau_b]) \leq bP([M^* > 0]) \leq (b-a)P([\tau_a \leq \tau_b]).$$

In words, $P([\tau_b \leq \tau_a])$ is the probability that $M(t)$ hits b no later than when it hits a. (Note that if $\tau_a = \infty = \tau_b$ then you would have $[\tau_a = \tau_b]$.)

Proof: For $x \in \mathbb{R}$, define

$$\tau_x \equiv \inf\{t \in \mathbb{R} \text{ such that } M(t) = x\}$$

with the usual convention that $\inf(\emptyset) = \infty$. Let $a < 0 < b$ and let

$$\tau = \tau_a \wedge \tau_b$$

Then the following claim will be important.

Claim: $E(M(\tau)) = 0$.

Proof of the claim: Let $t > 0$. Then by the Doob optional sampling theorem,

$$\begin{aligned} E(M(\tau \wedge t)) &= E(E(M(t)|\mathscr{F}_\tau)) = E(M(t)) && (31.8)\\ &= E(E(M(t)|\mathscr{F}_0)) = E(M(0)) = 0. && (31.9)\end{aligned}$$

Observe the martingale M^τ must be bounded because it is stopped when $M(t)$ equals either a or b. There are two cases according to whether $\tau = \infty$. If $\tau = \infty$, then $M(t)$ never hits a or b so $M(t)$ has values between a and b. In this case $M^\tau(t) = M(t) \in [a,b]$. On the other hand, you could have $\tau < \infty$. Then in this case $M^\tau(t)$ is eventually equal to either a or b depending on which it hits first. In either case, the martingale M^τ is bounded and by the martingale convergence theorem, Theorem 31.5.3, there exists $M^\tau(\infty)$ such that

$$\lim_{t\to\infty} M^\tau(t)(\omega) = M^\tau(\infty)(\omega) = M(\tau)(\omega)$$

and since the $M^\tau(t)$ are bounded, the dominated convergence theorem implies

$$E(M(\tau)) = \lim_{t\to\infty} E(M(\tau \wedge t)) = 0.$$

This proves the claim.

Let

$$M^*(\omega) \equiv \sup\{|M(t)(\omega)| : t \in [0,\infty]\}.$$

Also note that $[\tau_a = \tau_b] = [\tau = \infty]$. This is because $a \neq b$. If $M(t) = a$, then $M(t) \neq b$ so it cannot happen that these are equal at any finite time. But if $\tau = \infty$, then both $\tau_a, \tau_b = \infty$. Now from the claim,

$$\begin{aligned} 0 = E(M(\tau)) &= \int_{[\tau_a<\tau_b]} M(\tau)\,dP \\ &+ \int_{[\tau_b<\tau_a]} M(\tau)\,dP + \int_{[\tau_a=\tau_b]\cap[M^*>0]} M(\infty)\,dP \qquad (31.10)\\ &+ \int_{[\tau_a=\tau_b]\cap[M^*=0]} M(\infty)\,dP \end{aligned}$$

The last term equals 0. By continuity, $M(\tau)$ is either equal to a or b depending on whether $\tau_a < \tau_b$ or $\tau_b < \tau_a$. Thus

$$0 = E(M(\tau)) = aP([\tau_a < \tau_b])$$

$$+bP([\tau_b < \tau_a]) + \int_{[\tau_a=\tau_b]\cap[M^*>0]} M(\infty)\,dP \qquad (31.11)$$

Consider this last term. By the definition, $[\tau_a = \tau_b]$ corresponds to $M(t)$ never hitting either a or b. Since $M(0) = 0$, this can only happen if $M(t)$ has values in $[a,b]$. Therefore, this last term satisfies

$$\begin{aligned} aP([\tau_a = \tau_b] \cap [M^* > 0]) &\leq \int_{[\tau_a=\tau_b]\cap[M^*>0]} M(\infty)\, dP \\ &\leq bP([\tau_a = \tau_b] \cap [M^* > 0]) \end{aligned} \tag{31.12}$$

Obviously the following inequality holds because on the left you have

$$aP([\tau_a = \tau_b] \cap [M^* > 0])$$

and on the right you have the larger $bP([\tau_a = \tau_b] \cap [M^* > 0])$. That 0 is in the middle follows from 31.11.

$$aP([\tau_a = \tau_b] \cap [M^* > 0]) + aP([\tau_a < \tau_b]) + bP([\tau_b < \tau_a]) \leq$$

$$0 \leq bP([\tau_a = \tau_b] \cap [M^* > 0]) + aP([\tau_a < \tau_b]) + bP([\tau_b < \tau_a]) \tag{31.13}$$

Note that $[\tau_b < \tau_a], [\tau_a < \tau_b] \subseteq [M^* > 0]$ and so

$$[\tau_b < \tau_a] \cup [\tau_a < \tau_b] \cup ([\tau_a = \tau_b] \cap [M^* > 0]) = [M^* > 0] \tag{31.14}$$

The following diagram may help in keeping track of the various substitutions.

$$[M^* > 0]$$

$[\tau_a < \tau_b]$	$[\tau_b < \tau_a]$	$[\tau_b = \tau_a] \cap [M^* > 0]$

Left side of 31.13

From 31.14, this yields on substituting for $P([\tau_a < \tau_b])$

$$\begin{aligned} 0 \geq\ & aP([\tau_a = \tau_b] \cap [M^* > 0]) + a[P([M^* > 0]) - P([\tau_a \geq \tau_b] \cap [M^* > 0])] \\ & + bP([\tau_b < \tau_a]) \end{aligned}$$

and so since $[\tau_a \neq \tau_b] \subseteq [M^* > 0]$,

$$0 \geq a[P([M^* > 0]) - P([\tau_a > \tau_b])] + bP([\tau_b < \tau_a])$$

$$\boxed{-aP([M^* > 0]) \geq (b-a)P([\tau_b < \tau_a])} \tag{31.15}$$

Next use 31.14 to substitute for $P([\tau_b < \tau_a])$

$$0 \geq aP([\tau_a = \tau_b] \cap [M^* > 0]) + aP([\tau_a < \tau_b]) + bP([\tau_b < \tau_a])$$

$$\begin{aligned} &= aP([\tau_a = \tau_b] \cap [M^* > 0]) + aP([\tau_a < \tau_b]) \\ &\quad + b[P([M^* > 0]) - P([\tau_a \leq \tau_b] \cap [M^* > 0])] \end{aligned}$$

$$= aP([\tau_a \leq \tau_b] \cap [M^* > 0]) + b[P([M^* > 0]) - P([\tau_a \leq \tau_b] \cap [M^* > 0])]$$

and so

$$\boxed{(b-a)P([\tau_a \leq \tau_b]) \geq bP([M^* > 0])} \tag{31.16}$$

Right side of 31.13

From 31.14, used to substitute for $P([\tau_a < \tau_b])$ this yields

$$\begin{aligned} 0 \leq & bP([\tau_a = \tau_b] \cap [M^* > 0]) + aP([\tau_a < \tau_b]) + bP([\tau_b < \tau_a]) \\ = & bP([\tau_a = \tau_b] \cap [M^* > 0]) + a[P([M^* > 0]) - P([\tau_a \geq \tau_b] \cap [M^* > 0])] \\ & + bP([\tau_b < \tau_a]) \\ = & bP([\tau_a \geq \tau_b] \cap [M^* > 0]) + a[P([M^* > 0]) - P([\tau_a \geq \tau_b] \cap [M^* > 0])] \end{aligned}$$

and so

$$\boxed{(b-a)P([\tau_a \geq \tau_b]) \geq -aP([M^* > 0])} \tag{31.17}$$

Next use 31.14 to substitute for the term $P([\tau_b < \tau_a])$ and write

$$\begin{aligned} 0 \leq & bP([\tau_a = \tau_b] \cap [M^* > 0]) + aP([\tau_a < \tau_b]) + bP([\tau_b < \tau_a]) \\ = & bP([\tau_a = \tau_b] \cap [M^* > 0]) + aP([\tau_a < \tau_b]) \\ & + b[P([M^* > 0]) - P([\tau_a \leq \tau_b] \cap [M^* > 0])] \\ = & aP([\tau_a < \tau_b]) + bP([M^* > 0]) - bP([\tau_a < \tau_b] \cap [M^* > 0]) \\ = & aP([\tau_a < \tau_b]) + bP([M^* > 0]) - bP([\tau_a < \tau_b]) \end{aligned}$$

and so

$$\boxed{(b-a)P([\tau_a < \tau_b]) \leq bP([M^* > 0])} \tag{31.18}$$

Now the boxed in formulas in 31.15 - 31.18 yield the conclusion of the theorem. ■

Note $P([\tau_a < \tau_b])$ means $M(t)$ hits a before it hits b with other occurrences of similar expressions being defined similarly.

31.7 The Space $\mathscr{M}_T^p(E)$

Here $p \geq 1$. Also, we assume the filtration is a normal filtration.

Definition 31.7.1 *Then $M \in \mathscr{M}_T^p(E)$ if $t \to M(t)(\omega)$ is continuous for a.e. ω and $M(t)$ is adapted, and*

$$E\left(\sup_{t\in[0,T]} \|M(t)\|^p\right) < \infty$$

Here E is a separable Banach space.

Proposition 31.7.2 *Define a norm on $\mathscr{M}_T^p(E)$ by*

$$\|M\|_{\mathscr{M}_T^p(E)} \equiv E\left(\sup_{t\in[0,T]} \|M(t)\|^p\right)^{1/p}.$$

Then with this norm, $\mathscr{M}_T^p(E)$ is a Banach space. Also, a Cauchy sequence in this space has a subsequence which converges uniformly for all ω off a set of measure zero. Those M in $\mathscr{M}_T^p(E)$ which are martingales constitute a closed subspace of $\mathscr{M}_T^p(E)$. If σ is a stopping time, then if $M \in \mathscr{M}_T^p(E)$, so is M^σ and $\|M\|_{\mathscr{M}_T^p(E)} \geq \|M^\sigma\|_{\mathscr{M}_T^p(E)}$. Thus if $M_n \to M$ in $\mathscr{M}_T^p(E)$, then $M_n^\sigma \to M^\sigma$ in $\mathscr{M}_T^p(E)$.

Proof: First it is good to observe that $\sup_{t\in[0,T]}\|M(t)\|^p$ is measurable. This follows because of the continuity of $t\to M(t)$. Let D be a dense countable set in $[0,T]$. Then by continuity,

$$\sup_{t\in[0,T]}\|M(t)\|^p=\sup_{t\in D}\|M(t)\|^p$$

and the expression on the right is measurable because D is countable.

Next it is necessary to show this is a norm. It is clear that $\|M\|_{\mathscr{M}_T^p(E)}\geq 0$ and equals 0 only if $0=E\left(\sup_{t\in[0,T]}\|M(t)\|^p\right)$ which requires $M(t)=0$ for all t for ω off a set of measure zero so that $M=0$. It is also clear that $\|\alpha M\|_{\mathscr{M}_T^p(E)}=|\alpha|\,\|M\|_{\mathscr{M}_T^p(E)}$. It remains to check the triangle inequality. Let $M,N\in\mathscr{M}_T^p(E)$.

$$\|M+N\|_{\mathscr{M}_T^p(E)}\equiv E\left(\sup_{t\in[0,T]}\|M(t)+N(t)\|^p\right)^{1/p}$$

$$\begin{aligned}
&\leq E\left(\sup_{t\in[0,T]}(\|M(t)\|+\|N(t)\|)^p\right)^{1/p}\\
&\leq E\left(\left(\sup_{t\in[0,T]}\|M(t)\|+\sup_{t\in[0,T]}\|N(t)\|\right)^p\right)^{1/p}\\
&\equiv \left(\int_\Omega\left(\sup_{t\in[0,T]}\|M(t)\|+\sup_{t\in[0,T]}\|N(t)\|\right)^p dP\right)^{1/p}\\
&\leq \left(\int_\Omega\left(\sup_{t\in[0,T]}\|M(t)\|\right)^p dP\right)^{1/p}+\left(\int_\Omega\left(\sup_{t\in[0,T]}\|N(t)\|\right)^p dP\right)^{1/p}\\
&\equiv \|M\|_{\mathscr{M}_T^p(E)}+\|N\|_{\mathscr{M}_T^p(E)}
\end{aligned}$$

Next consider the claim that $\mathscr{M}_T^p(E)$ is a Banach space. Let $\{M_n\}$ be a Cauchy sequence. Then

$$E\left(\sup_{t\in[0,T]}\|M_n(t)-M_m(t)\|^p\right)\to 0 \tag{31.19}$$

as $m,n\to\infty$. Now

$$P\left(\sup_{t\in[0,T]}\|M_n(t)-M_m(t)\|>\lambda\right)\leq\frac{1}{\lambda^p}E\left(\sup_{t\in[0,T]}\|M_n(t)-M_m(t)\|\right)^p$$

Therefore, one can extract a subsequence $\{M_{n_k}\}$ such that

$$E\left(\sup_{t\in[0,T]}\left\|M_{n_k}(t)-M_{n_{k+1}}(t)\right\|^p\right)\leq 4^{-k}.$$

Then for this subsequence,

$$\begin{aligned} & P\left(\sup_{t\in[0,T]}\left\|M_{n_k}(t)-M_{n_{k+1}}(t)\right\|>2^{-k}\right) \\ \leq\ & 2^k E\left(\sup_{t\in[0,T]}\left\|M_{n_k}(t)-M_{n_{k+1}}(t)\right\|^p\right)\leq 2^{-k} \end{aligned}$$

and so, there is a set of measure zero N such that if $\omega\notin N$, then for all k large enough, $\sup_{t\in[0,T]}\left\|M_{n_k}(t)-M_{n_{k+1}}(t)\right\|\leq 2^{-k}$ and so for $\omega\notin N$, there exists M continuous such that $M_{n_k}(t)\to M(t)$ uniformly in $t\in[0,T]$. Thus for each ω, $\|M(t)\|^p\leq\left\|M_{n_k}(t)\right\|^p+\varepsilon$ for all $t\in[0,T]$ if k is large enough. Therefore, $\|M(t)\|^p\leq\sup_t\left\|M_{n_k}(t)\right\|^p+\varepsilon$ and so $\sup_t\|M(t)\|^p\leq\sup_t\left\|M_{n_k}(t)\right\|^p+\varepsilon$ for all t large enough. It follows that for each ω off a set of measure zero, $\sup_t\|M(t)\|^p\leq\liminf_{k\to\infty}\sup_t\left\|M_{n_k}(t)\right\|^p$. Is $M\in\mathscr{M}_T^p$? By Fatou's lemma,

$$\int_\Omega\sup_{t\in[0,T]}\|M(t)\|^p\,dP\leq\liminf_{k\to\infty}\left\|M_{n_k}\right\|^p_{\mathscr{M}_T^p(E)}$$

which is finite because $\{M_n\}$ is a Cauchy sequence. Thus $M\in\mathscr{M}_T^p(E)$. Now also,

$$\begin{aligned} & \left(\int_\Omega\sup_t\left\|M(t)-M_{n_k}(t)\right\|^p dP\right)^{1/p} \\ \leq\ & \liminf_{m\to\infty}\left(\int_\Omega\sup_t\left\|M_{n_m}(t)-M_{n_k}(t)\right\|^p dP\right)^{1/p}<\varepsilon \end{aligned}$$

if k is large enough because for $m>k$,

$$\begin{aligned} \left\|M_{n_m}-M_{n_k}\right\|_{\mathscr{M}_T^p} &\leq \sum_{r=k}^{\infty}\left\|M_{n_{r+1}}-M_{n_r}\right\|_{\mathscr{M}_T^p} \\ &\leq \sum_{r=k}^{\infty}\left(4^{1/p}\right)^{-r}=\left(\sqrt[p]{4}\right)^{-k}/\left(1-\left(\sqrt[p]{4}^{-1}\right)\right). \end{aligned}$$

This shows that every Cauchy sequence has a convergent subsequence and so the original Cauchy sequence also converges. This shows $\mathscr{M}_T^p$ is complete.

It only remains to verify that if each M_n is a martingale, then so is M a martingale. Let $s\leq t$ and let $B\in\mathscr{F}_s$. For each s, $M_n(s)\to M(s)$ in $L^p(\Omega)$. Then from the above, $\omega\to M(s)(\omega)$ is $\mathscr{F}_s$ measurable. Then it follows that

$$\begin{aligned} \int_B M(s)\,dP &= \lim_{n\to\infty}\int_B M_n(s)\,dP=\lim_{n\to\infty}\int_B E(M_n(t)\,|\mathscr{F}_s)\,dP \\ &= \lim_{n\to\infty}\int_B M_n(t)\,dP=\int_B M(t)\,dP \end{aligned}$$

and so by definition, $E(M(t)\,|\mathscr{F}_s)=M(s)$ which shows M is a martingale.

It is clear that if σ is a stopping time, then if $M\in\mathscr{M}_T^p(E)$ so is M^σ and that

$$\|M^\sigma\|_{\mathscr{M}_T^p(E)}\leq\|M\|_{\mathscr{M}_T^p(E)}$$

Thus if $M_n\to M$ in $\mathscr{M}_T^p(E)$, then $M_n^\sigma\to M^\sigma$ in $\mathscr{M}_T^p(E)$. ■

Note that if $M_n\to M$ in $\mathscr{M}_T^p(E)$, this says $\int_\Omega\sup_{t\in[0,T]}\|M_n(t)-M(t)\|^p\,dP=0$. Hence this would also be true that $\lim_{n\to\infty}\int_A\sup_{t\in[0,T]}\|M_n(t)-M(t)\|^p\,dP=0$ also, whenever A is a measurable set.

Chapter 32

The Quadratic Variation of a Martingale

32.1 How to Recognize a Martingale

The main ideas are most easily understood in the special case where it is assumed the martingale is bounded. Then one can extend to more general situations using a localizing sequence of stopping times.

Let $\{M(t)\}$ be a continuous martingale having values in a separable Hilbert space. The idea is to consider the submartingale, $\left\{\|M(t)\|^2\right\}$ and write it as the sum of a martingale and an increasing submartingale. An important part of the argument is the following lemma which gives a checkable criterion for a stochastic process to be a martingale.

Lemma 32.1.1 *Let $\{X(t)\}$ be a stochastic process adapted to the filtration $\{\mathscr{F}_t\}$ for $t \geq 0$. Then it is a martingale for the given filtration if for every stopping time σ it follows*

$$E(X(t)) = E(X(\sigma)).$$

In fact, it suffices to check this on stopping times which have two values.

Proof: Let $s < t$ and $A \in \mathscr{F}_s$. Define a stopping time

$$\sigma(\omega) \equiv s\mathscr{X}_A(\omega) + t\mathscr{X}_{A^C}(\omega)$$

This is a stopping time because $[\sigma \leq l] = \Omega \in \mathscr{F}_l$ if $l \geq t$. Also $[\sigma \leq l] = A \in \mathscr{F}_s \subseteq \mathscr{F}_l$ if $l \in [s,t)$ and $[\sigma \leq l] = \emptyset \in \mathscr{F}_l$ if $l < s$. Then by assumption,

$$\int_A X(t)\,dP + \int_{A^C} X(t)\,dP =$$

$$\overbrace{\int X(t)\,dP = \int X(\sigma)\,dP}^{\text{by assumption}} = \int_A X(s)\,dP + \int_{A^C} X(t)\,dP$$

Therefore,

$$\int_A X(t)\,dP = \int_A X(s)\,dP$$

and since $X(s)$ is $\mathscr{F}_s$ measurable, it follows $E(X(t)|\mathscr{F}_s) = X(s)$ a.e. and this shows $\{X(t)\}$ is a martingale. ■

Note that if $t \in [0,T]$, it suffices to check the expectation condition for stopping times which have two values no larger than T.

The following lemma will be useful.

Lemma 32.1.2 *Suppose $X_n \to X$ in $L^1(\Omega,\mathscr{F},P;E)$ where E is a separable Banach space. Then letting $\mathscr{G}$ be a σ algebra contained in $\mathscr{F}$,*

$$E(X_n|\mathscr{G}) \to E(X|\mathscr{G})$$

in $L^1(\Omega)$.

Proof: This follows from the definitions and Theorem 24.12.1 on Page 729.

$$\begin{aligned}\int_\Omega \|E(X|\mathscr{G}) - E(X_n|\mathscr{G})\|\,dP &= \int_\Omega \|E(X_n - X|\mathscr{G})\|\,dP \\ &\leq \int_\Omega E(\|X_n - X\||\mathscr{G})\,dP = \int_\Omega \|X_n - X\|\,dP \blacksquare\end{aligned}$$

The next corollary is like the earlier result which allows you to take a sufficiently measurable function out of the conditional expectation.

Corollary 32.1.3 *Let X, Y be in $L^2(\Omega,\mathscr{F},P;H)$ where H is a separable Hilbert space and let X be $\mathscr{G}$ measurable where $\mathscr{G}\subseteq\mathscr{F}$. Then*

$$E((X,Y)|\mathscr{G}) = (X, E(Y|\mathscr{G}))\ \text{a.e.}$$

Proof: First let $X = a\mathscr{X}_B$ where $B\in\mathscr{G}$. Then for $A\in\mathscr{G}$,

$$\begin{aligned}\int_A E((a\mathscr{X}_B,Y)|\mathscr{G})\,dP &= \int_A \mathscr{X}_B E((a,Y)|\mathscr{G})\,dP = \int_A \mathscr{X}_B(a,Y)\,dP \\ &= \int_{A\cap B}(a,Y)\,dP = \left(a, \int_{A\cap B} Y\,dP\right)\end{aligned}$$

$$\begin{aligned}\int_A (a\mathscr{X}_B, E(Y|\mathscr{G}))\,dP &= \int_A \mathscr{X}_B(a, E(Y|\mathscr{G}))\,dP \\ &= \left(a, \int_A \mathscr{X}_B E(Y|\mathscr{G})\,dP\right) = \left(a, \int_{A\cap B} Y\,dP\right)\end{aligned}$$

It follows that the formula holds for X simple.

Therefore, letting X_n be a sequence of $\mathscr{G}$ measurable simple functions converging pointwise to X and also in $L^2(\Omega)$,

$$E((X_n,Y)|\mathscr{G}) = (X_n, E(Y|\mathscr{G}))$$

Now the desired formula holds from Lemma 32.1.2. $\blacksquare$

The following is related to something called a martingale transform. It is a lot like what will happen later with the Ito integral.

Maybe it is a good idea to try and give some reason for considering the following. Say you have a bounded variation and adapted function f and you wanted to consider the Stieltjes integral $\int_0^T f\,dM$. If f is of bounded variation, this Stieltjes integral will exist from the standard theory of Stieltjes integration. In particular, if g is Stieltjes integrable with respect to df then f is Stieltjes integrable with respect to dg and an integration by parts formula holds. Now assuming M is continuous and f is of bounded variation, you would have the existence of $\int_0^T M\,df$ and so also the existence of $\int_0^T f\,dM$. Of course you might have different partitions for each different ω. In the following, this is handled by writing a sum of the form

$$\sum_{k\geq 0}(\xi_k, (M(\tau_{k+1}\wedge t) - M(\tau_k\wedge t)))$$

where ξ_k is in $\mathscr{F}_{\tau_k}$ and τ_k is a stopping time, the τ_k being an increasing sequence of stopping times having limit ∞. You could think of this as the value of f at the left end point. Of course what is happening here pertains to Hilbert space, but the inner product is

sufficiently like multiplication to draw the analogy and you would still have an integration by parts formula and the same result on existence of the integral. The theory of Stieltjes integrals is in my single variable advanced calculus book. An early reference to this is Hobson [28]. See also [2]. What is going to happen here is that these Stieltjes sums will end up being a martingale. The following is stated for the more general situation where $M(t)$ is only right continuous.

Proposition 32.1.4 *Let $\{\tau_k\}$ be an increasing sequence of stopping times for the normal filtration $\{\mathcal{F}_t\}$ such that*

$$\lim_{k\to\infty} \tau_k = \infty,\ \tau_0 = 0.$$

Also let ξ_k be $\mathcal{F}_{\tau_k}$ measurable with values in H, a separable Hilbert space and let $M(t)$ be a right continuous martingale adapted to the normal filtration $\mathcal{F}_t$ which has the property that $M(t) \in L^2(\Omega;H)$ for all $t, M(0) = 0$. Then if $|\xi_k| \le C$,

$$E\left(\left(\sum_{k\ge 0} (\xi_k, (M(\tau_{k+1}\wedge t) - M(\tau_k \wedge t)))\right)^2\right) \le C^2 E\left(\|M(t)\|^2\right) \tag{32.1}$$

Proof: First of all, the sum converges because eventually $\tau_k \wedge t = t$. Therefore, for large enough k, $M(\tau_{k+1}\wedge t) - M(\tau_k \wedge t) \equiv \Delta M_k = 0$. Consider first the finite sum, $k \le q$.

$$E\left(\left(\sum_{k=0}^{q} (\xi_k, \Delta M_k)\right)^2\right) \tag{32.2}$$

When the sum is multiplied out, you get mixed terms. Consider one of these mixed terms, $j < k$

$$E\left((\xi_k, \Delta M_k)\left(\xi_j, \Delta M_j\right)\right)$$

Using Corollary 32.1.3 and Doob's optional sampling theorem, this equals

$$E\left(E\left((\xi_k, \Delta M_k)\left(\xi_j, \Delta M_j\right)|\mathcal{F}_{\tau_k}\right)\right) = E\left(\left(\xi_j, \Delta M_j\right) E\left((\xi_k, \Delta M_k)|\mathcal{F}_{\tau_k}\right)\right)$$

$$\begin{aligned} &= E\left(\left(\xi_j, \Delta M_j\right)\left(\xi_k, E\left(M(\tau_{k+1}\wedge t) - M(\tau_k \wedge t)|\mathcal{F}_{\tau_k}\right)\right)\right) \\ &= E\left(\left(\xi_j, \Delta M_j\right)(\xi_k, 0)\right) = 0 \end{aligned}$$

Note that in using the optional sampling theorem, the stopping time $\tau_{k+1}\wedge t$ is bounded.

Therefore, the only terms which survive in 32.2 are the non mixed terms and so this expression reduces to

$$\sum_{k=0}^{q} E(\xi_k, \Delta M_k)^2 \le C^2 \sum_{k=0}^{q} E\left(\|\Delta M_k\|^2\right)$$

$$= C^2 \sum_{k=0}^{q} E\left(\|M(\tau_{k+1}\wedge t) - M(\tau_k \wedge t)\|^2\right)$$

$$\begin{aligned} &= C^2 \sum_{k=0}^{q} E\left(\|M(\tau_{k+1}\wedge t)\|^2\right) + E\left(\|M(\tau_k \wedge t)\|^2\right) \\ &\quad -2E\left(\left(M(\tau_k\wedge t), M(\tau_{k+1}\wedge t)\right)\right) \end{aligned} \tag{32.3}$$

Consider the term $E\left(\left(M(\tau_k\wedge t), M(\tau_{k+1}\wedge t)\right)\right)$. By Doob's optional sampling theorem for martingales and Corollary 32.1.3 again, this equals

$$\begin{aligned} & E\left(E\left(\left(M(\tau_k\wedge t), M(\tau_{k+1}\wedge t)\right)|\mathscr{F}_{\tau_k}\right)\right) \\ = \; & E\left(\left(M(\tau_k\wedge t), E\left(M(\tau_{k+1}\wedge t)|\mathscr{F}_{\tau_k}\right)\right)\right) \\ = \; & E\left(\left(M(\tau_k\wedge t), M(\tau_{k+1}\wedge t\wedge\tau_k)\right)\right) = E\left(\|M(\tau_k\wedge t)\|^2\right) \end{aligned}$$

It follows 32.3 equals

$$\begin{aligned} & C^2 \sum_{k=0}^{q} E\left(\|M(\tau_{k+1}\wedge t)\|^2\right) - E\left(\|M(\tau_k\wedge t)\|^2\right) \\ \leq \; & C^2 E\left(\|M(t)\|^2\right). \end{aligned}$$

Then from Fatou's lemma,

$$E\left(\left(\sum_{k\geq 0}(\xi_k, (M(\tau_{k+1}\wedge t) - M(\tau_k\wedge t)))\right)^2\right) \leq$$

$$\begin{aligned} & \liminf_{q\to\infty} E\left(\left(\sum_{k=0}^{q}(\xi_k, (M(\tau_{k+1}\wedge t) - M(\tau_k\wedge t)))\right)^2\right) \\ \leq \; & C^2 E\left(\|M(t)\|^2\right) \blacksquare \end{aligned}$$

Now here is an interesting lemma which will be used to prove uniqueness in the main result.

32.2 Martingales and Total Variation

Lemma 32.2.1 *Let $\mathscr{F}_t$ be a normal filtration and let $A(t), B(t)$ be adapted to $\mathscr{F}_t$, continuous, and increasing with $A(0) = B(0) = 0$ and suppose $A(t) - B(t)$ is a martingale. Then $A(t) - B(t) = 0$ for all t.*

Proof: I shall show $A(l) = B(l)$ where l is arbitrary. Let $M(t)$ be the name of the martingale. Define a stopping time

$$\begin{aligned} \tau \; \equiv \; & \inf\{t > 0 : |M(t)| > C\} \wedge l \wedge \inf\{t > 0 : A(t) > C\} \\ & \wedge \inf\{t > 0 : B(t) > C\} \end{aligned}$$

where $\inf(\emptyset) \equiv \infty$ and denote the stopped martingale $M^\tau(t) \equiv M(t\wedge\tau)$. Then this is also a martingale with respect to the filtration $\mathscr{F}_t$ because by Doob's optional sampling theorem for martingales. Recall why this is: if $s < t$,

$$E\left(M^\tau(t)|\mathscr{F}_s\right) \equiv E\left(M(\tau\wedge t)|\mathscr{F}_s\right) = M(\tau\wedge t\wedge s) = M(\tau\wedge s) = M^\tau(s)$$

Note the bounded stopping time is $\tau \wedge t$ and the other one is $\sigma = s$ in this theorem. Then M^τ is a continuous martingale which is also uniformly bounded. It equals $A^\tau - B^\tau$. The stopping time ensures A^τ and B^τ are uniformly bounded by C. Thus all of $|M^\tau(t)|, B^\tau(t), A^\tau(t)$ are bounded by C on $[0,l]$. Now let $\mathscr{P}_n \equiv \{t_k\}_{k=1}^n$ be a uniform partition of $[0,l]$ and let $M^\tau(\mathscr{P}_n)$ denote

$$M^\tau(\mathscr{P}_n) \equiv \max\{|M^\tau(t_{i+1}) - M^\tau(t_i)|\}_{i=1}^n.$$

Then

$$E\left(M^\tau(l)^2\right) = E\left(\left(\sum_{k=0}^{n-1} M^\tau(t_{k+1}) - M^\tau(t_k)\right)^2\right)$$

Now consider a mixed term in the sum where $j < k$.

$$E\left(\left(M^\tau(t_{k+1}) - M^\tau(t_k)\right)\left(M^\tau\left(t_{j+1}\right) - M^\tau(t_j)\right)\right)$$

$$\begin{aligned}
&= E\left(E\left(\left(M^\tau(t_{k+1}) - M^\tau(t_k)\right)\left(M^\tau\left(t_{j+1}\right) - M^\tau(t_j)\right)|\mathscr{F}_{t_k}\right)\right)\\
&= E\left(\left(M^\tau\left(t_{j+1}\right) - M^\tau(t_j)\right)E\left(\left(M^\tau(t_{k+1}) - M^\tau(t_k)\right)|\mathscr{F}_{t_k}\right)\right)\\
&= E\left(\left(M^\tau\left(t_{j+1}\right) - M^\tau(t_j)\right)\left(M^\tau(t_k) - M^\tau(t_k)\right)\right) = 0
\end{aligned}$$

It follows

$$\begin{aligned}
E\left(M^\tau(l)^2\right) &= E\left(\sum_{k=0}^{n-1}\left(M^\tau(t_{k+1}) - M^\tau(t_k)\right)^2\right)\\
&\leq E\left(\sum_{k=0}^{n-1} M^\tau(\mathscr{P}_n)\left|M^\tau(t_{k+1}) - M^\tau(t_k)\right|\right)
\end{aligned}$$

$$\begin{aligned}
&\leq E\left(\sum_{k=0}^{n-1} M^\tau(\mathscr{P}_n)\left(|A^\tau(t_{k+1}) - A^\tau(t_k)| + |B^\tau(t_{k+1}) - B^\tau(t_k)|\right)\right)\\
&\leq E\left(M^\tau(\mathscr{P}_n)\sum_{k=0}^{n-1}\left(|A^\tau(t_{k+1}) - A^\tau(t_k)| + |B^\tau(t_{k+1}) - B^\tau(t_k)|\right)\right)\\
&\leq E\left(M^\tau(\mathscr{P}_n)\,2C\right)
\end{aligned}$$

the last step holding because A and B are increasing. Now letting $n \to \infty$, the right side converges to 0 by the dominated convergence theorem and $\lim_{n\to\infty} M^\tau(\mathscr{P}_n)(\omega) = 0$ because of continuity of M. Thus for $\tau = \tau_C$ given above, $M(l \wedge \tau_C) = 0$ a.e. Now let $C \in \mathbb{N}$ and let N_C be the exceptional set off which $M(l \wedge \tau_C) = 0$. Then letting N_l denote the union of all these exceptional sets for $C \in \mathbb{N}$, it is also a set of measure zero and for ω not in this set, $M(l \wedge \tau_C) = 0$ for all C. Since the martingale is continuous, it follows for each such ω, eventually $\tau_C > l$ and so $M(l) = 0$. Thus for $\omega \notin N_l, M(l)(\omega) = 0$. Now let $N = \cup_{l\in\mathbb{Q}\cap[0,\infty)} N_l$. Then for $\omega \notin N, M(l)(\omega) = 0$ for all $l \in \mathbb{Q}\cap[0,\infty)$ and so by continuity, this is true for all positive l. ■

Note this shows a continuous martingale is not of bounded variation unless it is a constant.

If you had a continuous bounded variation function $f(t)$, you might want to do something like $\int_0^T \frac{1}{2}(f'(t), f(t))\,dt = \int_0^T \frac{1}{2}(f(t), df) = |f(T)|^2$. We do this all the time when

we discuss curves. From what was just shown, however, it will not be possible to do this directly in the context of Stieltjes integrals. What happens is something else. We get something called the quadratic variation of the martingale M denoted by $[M]$ which is increasing and $\|M\|^2 = [M] + N$ where N is a martingale. It will make perfect sense to write $\int_0^t f(s)\, d[M](s)$. Thus N does not have finite total variation but $[M]$ does.

32.3 The Quadratic Variation

This section is on the quadratic variation of a martingale. Actually, you can also consider the quadratic variation of a local martingale which is more general. Therefore, this concept is defined first. We will generally assume $M(0) = 0$ since there is no real loss of generality in doing so. One can simply subtract $M(0)$ otherwise. What is about to be presented is called the quadratic variation because it considers the variation of $\|M(t)\|^2$ rather than $M(t)$ which, as just shown is not finite.

Definition 32.3.1 *Let $\{M(t)\}$ be adapted to the normal filtration $\mathscr{F}_t$ for $t > 0$. Then $\{M(t)\}$ is a local martingale (submartingale) if there exist stopping times τ_n increasing to infinity such that for each n, the process $M^{\tau_n}(t) \equiv M(t \wedge \tau_n)$ is a martingale (submartingale) with respect to the given filtration. The sequence of stopping times is called a localizing sequence. The martingale M^{τ_n} is called the stopped martingale . Exactly the same convention applies to a localized submartingale. When M is continuous,we can always assume that M^{τ_n} is a bounded martingale (submartingale) by taking the minimum of τ_n with the first hitting time of $n \in \mathbb{N}$ by $\|M\|$. I will use this observation whenever convenient. If this is done, then $\tau_n = \infty$ for n large enough.*

Observation 32.3.2 *If M is a local martingale (submartingale) and if σ is a stopping time, then M^σ is also a local martingale (submartingale).*

To see this, use the localizing sequence. Say M is a local submartingale. The case of a martingale is similar. Let $s \leq t$.

$$E\left((M^\sigma)^{\tau_n}(t)\,|\mathscr{F}_s\right) = E\left(M^{\tau_n}(\sigma \wedge t)\,|\mathscr{F}_s\right) \geq M^{\tau_n}(\sigma \wedge s) = (M^\sigma)^{\tau_n}(s).$$

By the optional sampling theorem.

Proposition 32.3.3 *If $M(t)$ is a continuous local martingale (submartingale) for a normal filtration as above, $M(0) = 0$, then there exists a localizing sequence τ_n such that for each n the stopped martingale(submartingale) M^{τ_n} is uniformly bounded. Also if M is a martingale, then M^τ is also a martingale (submartingale). If τ_n is an increasing sequence of stopping times such that $\lim_{n\to\infty} \tau_n = \infty$, and for each τ_n and real valued stopping time δ, there exists a function X of $\tau_n \wedge \delta$ such that $X(\tau_n \wedge \delta)$ is $\mathscr{F}_{\tau_n \wedge \delta}$ measurable, then $\lim_{n\to\infty} X(\tau_n \wedge \delta) \equiv X(\delta)$ exists for each ω and $X(\delta)$ is $\mathscr{F}_\delta$ measurable.*

Proof: First review the claim about M^τ being a martingale (submartingale) when M is. By optional sampling theorem,

$$E(M^\tau(t)\,|\mathscr{F}_s) = E(M(\tau \wedge t)\,|\mathscr{F}_s) = M(\tau \wedge t \wedge s) = M^\tau(s).$$

The case where M is a submartingale is similar.

Next suppose σ_n is a localizing sequence for the local martingale(submartingale) M. Then define

$$\eta_n \equiv \inf\{t > 0 : \|M(t)\| > n\}.$$

Therefore, by continuity of M, $\|M(\eta_n)\| \leq n$. Now consider $\tau_n \equiv \eta_n \wedge \sigma_n$. This is an increasing sequence of stopping times. By continuity of M, it must be the case that $\eta_n \to \infty$. Hence $\sigma_n \wedge \eta_n \to \infty$.

Finally, consider the last claim. Pick ω. Then $X(\tau_n(\omega) \wedge \delta(\omega))(\omega)$ is eventually constant as $n \to \infty$ because for all n large enough, $\tau_n(\omega) > \delta(\omega)$ and so this sequence of functions converges pointwise. That which it converges to, denoted by $X(\delta)$, is $\mathscr{F}_\delta$ measurable because each function $\omega \to X(\tau_n(\omega) \wedge \delta(\omega))(\omega)$ is $\mathscr{F}_{\delta \wedge \tau_n} \subseteq \mathscr{F}_\delta$ measurable. ■

Observation 32.3.4 *Suppose M is a local martingale and τ_n is a localizing sequence of stoppings times. Does $M^{\tau_n}(t)$ converge in probability to $M(t)$? $M^{\tau_k}(t) = M(t)$ at ω where $\tau_k(\omega) = \infty$ and so $[\|M^{\tau_k}(t) - M(t)\| > \varepsilon] \subseteq [\tau_k < \infty]$ and $P([\tau_k < \infty]) \to 0$ by assumption that τ_k is a localizing sequence.*

One can also give a generalization of Lemma 32.2.1 to conclude a local martingale must be constant or else they must fail to be of bounded variation.

Corollary 32.3.5 *Let $\mathscr{F}_t$ be a normal filtration and let $A(t), B(t)$ be adapted to $\mathscr{F}_t$, continuous, and increasing with $A(0) = B(0) = 0$ and suppose $A(t) - B(t) \equiv M(t)$ is a local martingale. Then $M(t) = A(t) - B(t) = 0$ a.e. for all t.*

Proof: Let $\{\tau_n\}$ be a localizing sequence for M. For given n, consider the martingale,

$$M^{\tau_n}(t) = A^{\tau_n}(t) - B^{\tau_n}(t)$$

Then from Lemma 32.2.1, it follows $M^{\tau_n}(t) = 0$ for all t for all $\omega \notin N_n$, a set of measure 0. Let $N = \cup_n N_n$. Then for $\omega \notin N$, $M(\tau_n(\omega) \wedge t)(\omega) = 0$. Let $n \to \infty$ to conclude that $M(t)(\omega) = 0$. Therefore, $M(t)(\omega) = 0$ for all t. ■

Recall Example 31.3.13 on Page 869. For convenience, here is a version of what it says.

Lemma 32.3.6 *Let $X(t)$ be continuous and adapted to a normal filtration $\mathscr{F}_t$ and let η be a stopping time. Then if K is a closed set,*

$$\tau \equiv \inf\{t > \eta : X(t) \in K\}$$

is also a stopping time.

Proof: First consider $Y(t) = X(t \vee \eta) - X(\eta)$. I claim that $Y(t)$ is adapted to $\mathscr{F}_t$. Consider U and open set and $[Y(t) \in U]$. Is it in $\mathscr{F}_t$? We know it is in $\mathscr{F}_{t \vee \eta}$. It equals

$$([Y(t) \in U] \cap [\eta \leq t]) \cup ([Y(t) \in U] \cap [\eta > t])$$

Consider the second of these sets. It equals

$$([X(\eta) - X(\eta) \in U] \cap [\eta > t])$$

If $0 \in U$, then it reduces to $[\eta > t] \in \mathscr{F}_t$. If $0 \notin U$, then it reduces to $\emptyset$ still in $\mathscr{F}_t$. Next consider the first set. It equals

$$\begin{aligned} & [X(t \vee \eta) - X(\eta) \in U] \cap [\eta \leq t] \\ = \; & [X(t \vee \eta) - X(\eta) \in U] \cap [t \vee \eta \leq t] \in \mathscr{F}_t \end{aligned}$$

from the definition of $\mathscr{F}_{t\vee\eta}$. (You know that $[X(t \vee \eta) - X(\eta) \in U] \in \mathscr{F}_{t\vee\eta}$ and so when this is intersected with $[t \vee \eta \leq t]$ one obtains a set in $\mathscr{F}_t$. This is what it means to be in $\mathscr{F}_{t\vee\eta}$.) Now τ is just the first hitting time of $Y(t)$ of the closed set. ■

Proposition 32.3.7 *Let $M(t)$ be a continuous local martingale for $t \in [0,T]$ having values in H a separable Hilbert space adapted to the normal filtration $\{\mathscr{F}_t\}$ such that $M(0) = 0$. Then there exists a unique continuous, increasing, nonnegative, local submartingale $[M](t)$ called the quadratic variation such that*

$$\|M(t)\|^2 - [M](t)$$

is a real local martingale and $[M](0) = 0$. Here $t \in [0,T]$. If δ is any stopping time

$$\left[M^\delta\right] = [M]^\delta$$

Proof: First it is necessary to define some stopping times. Define stopping times $\tau_0^n \equiv \eta_0^n \equiv 0$.

$$\begin{aligned} \eta_{k+1}^n &\equiv \inf\left\{s > \eta_k^n : \|M(s) - M(\eta_k^n)\| = 2^{-n}\right\}, \\ \tau_k^n &\equiv \eta_k^n \wedge T \end{aligned}$$

where $\inf\emptyset \equiv \infty$. These are stopping times by Example 31.3.13 on Page 869. See also the above Lemma 32.3.6. Then for $t > 0$ and δ any stopping time, and fixed ω, for some k,

$$t \wedge \delta \in I_k(\omega),\ I_0(\omega) \equiv [\tau_0^n(\omega), \tau_1^n(\omega)],\ I_k(\omega) \equiv (\tau_k^n(\omega), \tau_{k+1}^n(\omega)] \text{ some } k$$

Here is why. The sequence $\left\{\tau_k^n(\omega)\right\}_{k=1}^\infty$ eventually equals T for all n sufficiently large. This is because if it did not, it would converge, being bounded above by T and then by continuity of M, $\left\{M\left(\tau_k^n(\omega)\right)\right\}_{k=1}^\infty$ would be a Cauchy sequence contrary to the requirement that

$$\begin{aligned} & \left\|M\left(\tau_{k+1}^n(\omega)\right) - M\left(\tau_k^n(\omega)\right)\right\| \\ = \; & \left\|M\left(\eta_{k+1}^n(\omega)\right) - M\left(\eta_k^n(\omega)\right)\right\| = 2^{-n}. \end{aligned}$$

Note that if δ is any stopping time, then

$$\begin{aligned} & \left\|M\left(t \wedge \delta \wedge \tau_{k+1}^n\right) - M\left(t \wedge \delta \wedge \tau_k^n\right)\right\| \\ = \; & \left\|M^\delta\left(t \wedge \tau_{k+1}^n\right) - M^\delta\left(t \wedge \tau_k^n\right)\right\| \leq 2^{-n} \end{aligned}$$

You can see this is the case by considering the cases, $t \wedge \delta \geq \tau_{k+1}^n, t \wedge \delta \in [\tau_k^n, \tau_{k+1}^n)$, and $t \wedge \delta < \tau_k^n$. It is only this approximation property and the fact that the τ_k^n partition $[0,T]$ which is important in the following argument.

Now let α_n be a localizing sequence such that M^{α_n} is bounded as in Proposition 32.3.3. Thus $M^{\alpha_n}(t) \in L^2(\Omega)$ and this is all that is needed. In what follows, let δ be a stopping

time and denote $M^{\alpha_p \wedge \delta}$ by M to save notation. Thus M will be uniformly bounded and from the definition of the stopping times τ_k^n, for $t \in [0,T]$,

$$M(t) \equiv \sum_{k\geq 0} M\left(t\wedge\tau_{k+1}^n\right) - M\left(t\wedge\tau_k^n\right), \tag{32.4}$$

and the terms of the series are eventually 0, as soon as $\tau_k^n = \infty$.

Therefore,

$$\|M(t)\|^2 = \left\| \sum_{k\geq 0} M\left(t\wedge\tau_{k+1}^n\right) - M\left(t\wedge\tau_k^n\right) \right\|^2$$

Then this equals

$$= \sum_{k\geq 0} \left\| M\left(t\wedge\tau_{k+1}^n\right) - M\left(t\wedge\tau_k^n\right) \right\|^2$$

$$+ \sum_{j\neq k} \left(\left(M\left(t\wedge\tau_{k+1}^n\right) - M\left(t\wedge\tau_k^n\right) \right), \left(M\left(t\wedge\tau_{j+1}^n\right) - M\left(t\wedge\tau_j^n\right) \right) \right) \tag{32.5}$$

Consider the second sum. It equals

$$2\sum_{k\geq 0}\sum_{j=0}^{k-1} \left(\left(M\left(t\wedge\tau_{k+1}^n\right) - M\left(t\wedge\tau_k^n\right) \right), \left(M\left(t\wedge\tau_{j+1}^n\right) - M\left(t\wedge\tau_j^n\right) \right) \right)$$

$$\begin{aligned} &= 2\sum_{k\geq 0} \left(\left(M\left(t\wedge\tau_{k+1}^n\right) - M\left(t\wedge\tau_k^n\right) \right), \overbrace{\sum_{j=0}^{k-1} \left(M\left(t\wedge\tau_{j+1}^n\right) - M\left(t\wedge\tau_j^n\right) \right)}^{\text{telescopes}} \right) \\ &= 2\sum_{k\geq 0} \left(\left(M\left(t\wedge\tau_{k+1}^n\right) - M\left(t\wedge\tau_k^n\right) \right), M\left(t\wedge\tau_k^n\right) \right) \end{aligned}$$

This last sum equals $P_n(t)$ defined as

$$2\sum_{k\geq 0} \left(M\left(\tau_k^n\right), \left(M\left(t\wedge\tau_{k+1}^n\right) - M\left(t\wedge\tau_k^n\right) \right) \right) \equiv P_n(t) \tag{32.6}$$

This is because in the k^{th} term, if $t \geq \tau_k^n$, then it reduces to

$$\left(M\left(\tau_k^n\right), \left(M\left(t\wedge\tau_{k+1}^n\right) - M\left(t\wedge\tau_k^n\right) \right) \right)$$

while if $t < \tau_k^n$, then the term reduces to $\left(\left(M(t) - M(t) \right), M(t) \right) = 0$ which is also the same as

$$\left(M\left(\tau_k^n\right), \left(M\left(t\wedge\tau_{k+1}^n\right) - M\left(t\wedge\tau_k^n\right) \right) \right).$$

This is a finite sum because eventually, for large enough k, $\tau_k^n = T$. However the number of nonzero terms depends on ω. This is not a good thing. However, a little more can be said. In fact the sum in 32.6 converges in $L^2(\Omega)$. Say $\|M(t,\omega)\| \leq C$.

$$E\left(\left(\sum_{k\geq p}^{q} \left(M\left(\tau_k^n\right), \left(M\left(t\wedge\tau_{k+1}^n\right) - M\left(t\wedge\tau_k^n\right) \right) \right) \right)^2 \right)$$

$$= \sum_{k\geq p}^{q} E\left(\left(M\left(\tau_k^n\right), \left(M\left(t\wedge\tau_{k+1}^n\right) - M\left(t\wedge\tau_k^n\right)\right)\right)^2\right) + \text{ mixed terms} \tag{32.7}$$

Consider one of these mixed terms for $j<k$.

$$E\left(\left(M\left(\tau_j^n\right), \left(\overbrace{M\left(t\wedge\tau_{j+1}^n\right) - M\left(t\wedge\tau_j^n\right)}^{\Delta_j}\right)\right)\cdot \left(M\left(\tau_k^n\right), \left(\overbrace{M\left(t\wedge\tau_{k+1}^n\right) - M\left(t\wedge\tau_k^n\right)}^{\Delta_k}\right)\right)\right)$$

Then it equals

$$\begin{aligned} & E\left(E\left(\left(M\left(\tau_j^n\right),\Delta_j\right)\left(M\left(\tau_j^n\right),\Delta_k\right)|\mathscr{F}_{\tau_k}\right)\right) \\ = \; & E\left(\left(M\left(\tau_j^n\right),\Delta_j\right)E\left(\left(M\left(\tau_j^n\right),\Delta_k\right)|\mathscr{F}_{\tau_k}\right)\right) \\ = \; & E\left(\left(M\left(\tau_j^n\right),\Delta_j\right)\left(M\left(\tau_j^n\right),E\left(\Delta_k|\mathscr{F}_{\tau_k}\right)\right)\right) = 0 \end{aligned}$$

since $E\left(\Delta_k|\mathscr{F}_{\tau_k}\right) = E\left(M\left(t\wedge\tau_{k+1}^n\right) - M\left(t\wedge\tau_k^n\right)|\mathscr{F}_{\tau_k}\right) = 0$. Now since the mixed terms equal 0, it follows from 32.7, that expression is dominated by

$$C^2 \sum_{k\geq p}^{q} E\left(\left\|M\left(t\wedge\tau_{k+1}^n\right) - M\left(t\wedge\tau_k^n\right)\right\|^2\right) \tag{32.8}$$

A mixed term in the above is of the form: For $j<k$,

$$E\left(\Delta_k,\Delta_j\right) = E\left(E\left(\left(\Delta_k,\Delta_j\right)|\mathscr{F}_{\tau_k^n}\right)\right) = E\left(\left(\Delta_j, E\left(\Delta_k|\mathscr{F}_{\tau_k^n}\right)\right)\right) = 0$$

Thus 32.8 equals

$$\begin{aligned} & C^2 \sum_{k=p}^{q} E\left(\left\|M\left(t\wedge\tau_{k+1}^n\right)\right\|^2\right) - E\left(\left\|M\left(t\wedge\tau_k^n\right)\right\|^2\right) \\ = \; & C^2 E\left(\left\|M\left(t\wedge\tau_{q+1}^n\right)\right\|^2 - \left\|M\left(t\wedge\tau_p^n\right)\right\|^2\right) \end{aligned}$$

The integrand converges to 0 as $p,q\to\infty$ and the uniform bound on M allows a use of the dominated convergence theorem. Thus the partial sums of the series of 32.6 converge in $L^2(\Omega)$ as claimed.

By adding in the values of $\left\{\tau_k^{n+1}\right\}$ $P_n(t)$ can be written in the form

$$2\sum_{k\geq 0}\left(M\left(\tau_k^{n+1\prime}\right), \left(M\left(t\wedge\tau_{k+1}^{n+1}\right) - M\left(t\wedge\tau_k^{n+1}\right)\right)\right)$$

where $\tau_k^{n+1\prime}$ has some repeats. From the construction,

$$\left\|M\left(\tau_k^{n+1\prime}\right) - M\left(\tau_k^{n+1}\right)\right\| \leq 2^{-(n+1)}$$

Thus

$$P_n(t) - P_{n+1}(t) = 2\sum_{k\geq 0}\left(M\left(\tau_k^{n+1\prime}\right) - M\left(\tau_k^{n+1}\right), \left(M\left(t\wedge\tau_{k+1}^{n+1}\right) - M\left(t\wedge\tau_k^{n+1}\right)\right)\right)$$

and so from Proposition 32.1.4 applied to $\xi_k \equiv M\left(\tau_k^{n+1\prime}\right) - M\left(\tau_k^{n+1}\right)$,

$$E\left(\|P_n(t) - P_{n+1}(t)\|^2\right) \leq \left(2^{-2n} E\left(\|M(t)\|^2\right)\right). \tag{32.9}$$

Now $t \to P_n(t)$ is continuous because it is a finite sum of continuous functions. It is also the case that $\{P_n(t)\}$ is a martingale. To see this use Lemma 32.1.1. Let σ be a stopping time having two values. Then using Corollary 32.1.3 and the Doob optional sampling theorem, Theorem 31.3.16

$$E\left(\sum_{k=0}^{q}\left(M\left(\tau_k^n\right), \left(M\left(\sigma \wedge \tau_{k+1}^n\right) - M\left(\sigma \wedge \tau_k^n\right)\right)\right)\right)$$

$$= \sum_{k=0}^{q} E\left(\left(M\left(\tau_k^n\right), \left(M\left(\sigma \wedge \tau_{k+1}^n\right) - M\left(\sigma \wedge \tau_k^n\right)\right)\right)\right)$$

$$\begin{aligned}
&= \sum_{k=0}^{q} E\left(\left(E\left(M\left(\tau_k^n\right), \left(M\left(\sigma \wedge \tau_{k+1}^n\right) - M\left(\sigma \wedge \tau_k^n\right)\right)\right) | \mathscr{F}_{\tau_k^n}\right)\right) \\
&= \sum_{k=0}^{q} E\left(\left(M\left(\tau_k^n\right), E\left(M\left(\sigma \wedge \tau_{k+1}^n\right) - M\left(\sigma \wedge \tau_k^n\right)\right) | \mathscr{F}_{\tau_k^n}\right)\right) \\
&= \sum_{k=0}^{q} E\left(\left(M\left(\tau_k^n\right), E\left(M\left(\sigma \wedge \tau_{k+1}^n \wedge \tau_k^n\right) - M\left(\sigma \wedge \tau_k^n\right)\right)\right)\right) = 0
\end{aligned}$$

Note the Doob theorem applies because $\sigma \wedge \tau_{k+1}^n$ is a bounded stopping time due to the fact σ has only two values. Similarly

$$E\left(\sum_{k=0}^{q}\left(M\left(\tau_k^n\right), \left(M\left(t \wedge \tau_{k+1}^n\right) - M\left(t \wedge \tau_k^n\right)\right)\right)\right)$$

$$= \sum_{k=0}^{q} E\left(\left(M\left(\tau_k^n\right), \left(M\left(t \wedge \tau_{k+1}^n\right) - M\left(t \wedge \tau_k^n\right)\right)\right)\right)$$

$$\begin{aligned}
&= \sum_{k=0}^{q} E\left(\left(E\left(M\left(\tau_k^n\right), \left(M\left(t \wedge \tau_{k+1}^n\right) - M\left(t \wedge \tau_k^n\right)\right)\right) | \mathscr{F}_{\tau_k^n}\right)\right) \\
&= \sum_{k=0}^{q} E\left(\left(M\left(\tau_k^n\right), E\left(M\left(t \wedge \tau_{k+1}^n\right) - M\left(t \wedge \tau_k^n\right)\right) | \mathscr{F}_{\tau_k^n}\right)\right) \\
&= \sum_{k=0}^{q} E\left(\left(M\left(\tau_k^n\right), E\left(M\left(t \wedge \tau_{k+1}^n \wedge \tau_k^n\right) - M\left(t \wedge \tau_k^n\right)\right)\right)\right) = 0
\end{aligned}$$

It follows each partial sum for $P_n(t)$ is a martingale. As shown above, these partial sums converge in $L^2(\Omega)$ and so it follows that $P_n(t)$ is also a martingale. Note the Doob theorem applies because $t \wedge \tau_{k+1}^n$ is a bounded stopping time.

I want to argue that P_n is a Cauchy sequence in $\mathscr{M}_T^2(\mathbb{R})$. By Theorem 31.4.3 and continuity of P_n which yields appropriate measurability in $\sup_{t \leq T} |P_n(t) - P_{n+1}(t)|$,

$$E\left(\left(\sup_{t \leq T} |P_n(t) - P_{n+1}(t)|\right)^2\right)^{1/2} \leq 2E\left(|P_n(T) - P_{n+1}(T)|^2\right)^{1/2}$$

By 32.9, $\leq 2^{-n} E\left(\|M(T)\|^2\right)^{1/2}$ which shows $\{P_n\}$ is a Cauchy sequence in $\mathscr{M}_T^2(\mathbb{R})$.

Therefore, by Proposition 31.7.2, there exists $\{N(t)\} \in \mathscr{M}_T^2(\mathbb{R})$ such that $P_n \to N$ in $\mathscr{M}_T^2(H)$. That is

$$\lim_{n\to\infty} E\left(\sup_{t\in[0,T]} |P_n(t) - N(t)|^2\right)^{1/2} = 0.$$

Since $\{N(t)\} \in \mathscr{M}_T^2(\mathbb{R})$, it is a continuous martingale and $N(t) \in L^2(\Omega)$, and $N(0) = 0$ because this is true of each $P_n(0)$. From the above 32.5,

$$\|M(t)\|^2 = Q_n(t) + P_n(t) \tag{32.10}$$

where

$$Q_n(t) = \sum_{k\geq 0} \left\|M\left(t\wedge\tau_{k+1}^n\right) - M\left(t\wedge\tau_k^n\right)\right\|^2$$

and $P_n(t)$ is a martingale. Then from 32.10, $Q_n(t)$ is a submartingale and converges for each t to something, denoted as $[M](t)$ in $L^1(\Omega)$ uniformly in $t \in [0,T]$. This is because $P_n(t)$ converges uniformly on $[0,T]$ to $N(t)$ in $L^2(\Omega)$ and $\|M(t)\|^2$ does not depend on n. Then also $[M]$ is a submartingale which equals 0 at 0 because this is true of Q_n and because if $A \in \mathscr{F}_s$ where $s < t$,

$$\int_A E([M](t)|\mathscr{F}_s)\,dP \equiv \int_A [M](t)\,dP = \lim_{n\to\infty}\int_A \left(\|M(t)\|^2 - P_n(t)\right)dP$$

$$= \lim_{n\to\infty}\int_A E\left(\|M(t)\|^2 - P_n(t)|\mathscr{F}_s\right)dP \geq \liminf_{n\to\infty}\int_A \|M(s)\|^2 - P_n(s)\,dP$$

$$= \liminf_{n\to\infty}\int_A Q_n(s)\,dP \geq \int_A [M](s)\,dP.$$

Note that $Q_n(t)$ is increasing because as t increases, the definition allows for the possibility of more nonzero terms in the sum. Therefore, $[M](t)$ is also increasing in t. The function $t \to [M](t)$ is continuous because $\|M(t)\|^2 = [M](t) + N(t)$ and $t \to N(t)$ is continuous as is $t \to \|M(t)\|^2$. That is, off a set of measure zero, these are both continuous functions of t and so the same is true of $[M]$.

Now put back in $M^{\alpha_p\wedge\delta}$ in place of M where δ is a stopping time. From the above, this has shown

$$\left\|M^{\alpha_p\wedge\delta}(t)\right\|^2 = \left[M^{\alpha_p\wedge\delta}\right](t) + N_p(t)$$

where N_p is a martingale and

$$\left[M^{\alpha_p\wedge\delta}\right](t) = \lim_{n\to\infty}\sum_{k\geq 0}\left\|M^{\alpha_p\wedge\delta}\left(t\wedge\tau_{k+1}^n\right) - M^{\alpha_p\wedge\delta}\left(t\wedge\tau_k^n\right)\right\|^2$$

$$= \lim_{n\to\infty}\sum_{k\geq 0}\left\|M\left(t\wedge\tau_{k+1}^n\wedge\alpha_p\wedge\delta\right) - M\left(t\wedge\tau_k^n\wedge\alpha_p\wedge\delta\right)\right\|^2 \text{ in } L^1(\Omega), \tag{32.11}$$

the convergence being uniform on $[0,T]$. The above formula shows that $\left[M^{\alpha_p\wedge\delta}\right](t)$ is a $\mathscr{F}_{t\wedge\delta\wedge\alpha_p}$ measurable random variable which depends on $t\wedge\delta\wedge\alpha_p$.(Note that $t\wedge\delta$ is a real valued stopping time even if $\delta = \infty$.) Therefore, by Proposition 32.3.3, there exists a random variable, denoted as $\left[M^{\delta}\right](t)$ which is the pointwise limit as $p \to \infty$ of these random

variables which is $\mathscr{F}_{t\wedge\delta}$ measurable because, for a given ω, when α_p becomes larger than t, the sum in 32.11 loses its dependence on p. Thus from pointwise convergence in 32.11,

$$\left[M^{\delta}\right](t) \equiv \lim_{n\to\infty} \sum_{k\geq 0} \left|M\left(t\wedge\delta\wedge\tau_{k+1}^{n}\right) - M\left(t\wedge\delta\wedge\tau_{k}^{n}\right)\right|^{2}$$

In case $\delta = \infty$, the above gives an $\mathscr{F}_t$ measurable random variable denoted by $[M](t)$ such that

$$[M](t) \equiv \lim_{n\to\infty} \sum_{k\geq 0} \left|M\left(t\wedge\tau_{k+1}^{n}\right) - M\left(t\wedge\tau_{k}^{n}\right)\right|^{2}$$

Now stopping with the stopping time δ, this shows that

$$\left[M^{\delta}\right](t) \equiv \lim_{n\to\infty} \sum_{k\geq 0} \left|M\left(t\wedge\delta\wedge\tau_{k+1}^{n}\right) - M\left(t\wedge\delta\wedge\tau_{k}^{n}\right)\right|^{2} = [M]^{\delta}(t)$$

That is, the quadratic variation of the stopped local martingale makes sense a.e. and equals the stopped quadratic variation of the local martingale.

This has now shown that

$$\begin{aligned} \|M^{\alpha_n}(t)\|^2 - [M]^{\alpha_n}(t) &= \|M^{\alpha_n}(t)\|^2 - [M^{\alpha_n}](t) \\ &= N_n(t),\ N_n(t) \text{ a martingale} \end{aligned}$$

and both of the random variables on the left converge pointwise as $n\to\infty$ to a function which is $\mathscr{F}_t$ measurable. Hence so does $N_n(t)$. Of course $N_n(t)$ is likewise a function of $\alpha_n\wedge t$ and so by Proposition 32.3.3 again, it converges pointwise to a $\mathscr{F}_t$ measurable function called $N(t)$ and $N(t)$ is a continuous local martingale.

It remains to consider the claim about the uniqueness. Suppose then there are two which work, $[M]$, and $[M]_1$. Then $[M]-[M]_1$ equals a local martingale G which is 0 when $t=0$. Thus the uniqueness assertion follows from Corollary 32.3.5. ■

Here is a corollary which tells how to manipulate stopping times. It is contained in the above proposition, but it is worth emphasizing it from a different point of view.

Corollary 32.3.8 *In the situation of Proposition 32.3.7 let τ be a stopping time. Then*

$$[M^{\tau}] = [M]^{\tau}.$$

Proof:

$$[M]^{\tau}(t) + N_1(t) = \left(\|M\|^2\right)^{\tau}(t) = \|M^{\tau}\|^2(t) = [M^{\tau}](t) + N_2(t)$$

where N_i is a local martingale. Therefore,

$$[M]^{\tau}(t) - [M^{\tau}](t) = N_2(t) - N_1(t),$$

a local martingale. Therefore, by Corollary 32.3.5, this shows $[M]^{\tau}(t) - [M^{\tau}](t) = 0$. ■

32.4 The Covariation

Definition 32.4.1 *The covariation of two continuous H valued local martingales for H a separable Hilbert space $M, N, M(0) = 0 = N(0)$, is defined as follows.*

$$[M,N] \equiv \frac{1}{4}\left([M+N] - [M-N]\right)$$

Lemma 32.4.2 *The following hold for the covariation.*

$$[M] = [M,M]$$

$$\begin{aligned}[M,N] &= \textit{local martingale} + \frac{1}{4}\left(\|M+N\|^2 - \|M-N\|^2\right) \\ &= (M,N) + \textit{local martingale.}\end{aligned}$$

Proof: From the definition of covariation,

$$[M] = \|M\|^2 - \mathscr{N}_1$$

$$\begin{aligned}[M,M] &= \frac{1}{4}\left([M+M]-[M-M]\right) = \frac{1}{4}\left(\|M+M\|^2 - \mathscr{N}_2\right) \\ &= \|M\|^2 - \frac{1}{4}\mathscr{N}_2\end{aligned}$$

where $\mathscr{N}_i$ is a local martingale. Thus $[M]-[M,M]$ is equal to the difference of two increasing continuous adapted processes and it also equals a local martingale. By Corollary 32.3.5, this process must equal 0. Now consider the second claim.

$$\begin{aligned}[M,N] &= \frac{1}{4}\left([M+N]-[M-N]\right) = \frac{1}{4}\left(\|M+N\|^2 - \|M-N\|^2 + \mathscr{N}\right) \\ &= (M,N) + \frac{1}{4}\mathscr{N}\end{aligned}$$

where $\mathscr{N}$ is a local martingale. ■

Corollary 32.4.3 *Let M,N be two continuous local martingales,*

$$M(0) = N(0) = 0,$$

as in Proposition 32.3.7. Then $[M,N]$ is of bounded variation and

$$(M,N)_H - [M,N]$$

is a local martingale. Also for τ a stopping time,

$$[M,N]^\tau = [M^\tau,N^\tau] = [M^\tau,N] = [M,N^\tau].$$

In addition to this,

$$[M-M^\tau] = [M] - [M^\tau] \leq [M]$$

and also

$$M,N \to [M,N]$$

is bilinear and symmetric.

Proof: Since $[M,N]$ is the difference of increasing functions, it is of bounded variation.

$$\begin{aligned}(M,N)_H - [M,N] &= \overbrace{\frac{1}{4}\left(\|M+N\|^2 - \|M-N\|^2\right)}^{(M,N)_H} \\ &\quad - \overbrace{\frac{1}{4}\left([M+N]-[M-N]\right)}^{[M,N]}\end{aligned}$$

which equals a local martingale from the definition of $[M+N]$ and $[M-N]$. It remains to verify the claim about the stopping time. Using Corollary 32.3.8

$$[M,N]^\tau = \frac{1}{4}\left([M+N]-[M-N]\right)^\tau$$

$$\begin{aligned} &= \frac{1}{4}\left([M+N]^\tau-[M-N]^\tau\right) \\ &= \frac{1}{4}\left([M^\tau+N^\tau]-[M^\tau-N^\tau]\right) \equiv [M^\tau,N^\tau]. \end{aligned}$$

The really interesting part is the next equality. This will involve Corollary 32.3.5.

$$[M,N]^\tau - [M^\tau,N] = [M^\tau,N^\tau] - [M^\tau,N]$$

$$\begin{aligned} &\equiv \frac{1}{4}\left([M^\tau+N^\tau]-[M^\tau-N^\tau]\right) - \frac{1}{4}\left([M^\tau+N]-[M^\tau-N]\right) \\ &= \frac{1}{4}\left([M^\tau+N^\tau]+[M^\tau-N]\right) - \frac{1}{4}\left([M^\tau+N]+[M^\tau-N^\tau]\right), \end{aligned} \tag{32.12}$$

the difference of two increasing adapted processes. Also, this equals

$$\text{local martingale } -(M^\tau,N)+(M^\tau,N^\tau)$$

Claim: $(M^\tau,N)-(M^\tau,N^\tau)=(M^\tau,N-N^\tau)$ is a local martingale. Let σ_n be a localizing sequence for both M and N. Such a localizing sequence is of the form $\tau_n^M \wedge \tau_n^N$ where these are localizing sequences for the indicated local submartingale. Then obviously,

$$\left(-(M^\tau,N)+(M^\tau,N^\tau)\right)^{\sigma_n} = -\left(M^{\sigma_n\wedge\tau},N^{\sigma_n}\right)+\left(M^{\sigma_n\wedge\tau},N^{\sigma_n\wedge\tau}\right)$$

where N^{σ_n} and M^{σ_n} are martingales. To save notation, denote these by M and N respectively. Now use Lemma 32.1.1. Let σ be a stopping time with two values.

$$E\left(\left(M^\tau(\sigma),N(\sigma)-N^\tau(\sigma)\right)\right) = E\left(E\left(\left(M^\tau(\sigma),N(\sigma)-N^\tau(\sigma)\right)|\mathcal{F}_\tau\right)\right)$$

Now $M^\tau(\sigma)$ is $M(\sigma\wedge\tau)$ which is $\mathcal{F}_\tau$ measurable and so by the Doob optional sampling theorem,

$$\begin{aligned} &= E\left(M^\tau(\sigma),E\left(N(\sigma)-N^\tau(\sigma)|\mathcal{F}_\tau\right)\right) \\ &= E\left(M^\tau(\sigma),N(\sigma\wedge\tau)-N(\tau\wedge\sigma)\right)=0 \end{aligned}$$

while

$$E\left(\left(M^\tau(t),N(t)-N^\tau(t)\right)\right) = E\left(E\left(\left(M^\tau(t),N(t)-N^\tau(t)\right)|\mathcal{F}_\tau\right)\right)$$

Since $M^\tau(t)$ is $\mathcal{F}_\tau$ measurable,

$$\begin{aligned} &= E\left(\left(M^\tau(t),E\left(N(t)-N^\tau(t)|\mathcal{F}_\tau\right)\right)\right) \\ &= E\left(\left(M^\tau(t),E\left(N(t\wedge\tau)-N(t\wedge\tau)\right)\right)\right)=0 \end{aligned}$$

This shows the claim is true.

Now from 32.12 and Corollary 32.4.3, $[M,N]^\tau - [M^\tau,N] = 0$. Similarly $[M,N]^\tau - [M,N^\tau] = 0$. Now consider the next claim that $[M-M^\tau] = [M] - [M^\tau]$. From the definition, it follows

$$\begin{aligned} & [M-M^\tau] - ([M] + [M^\tau] - 2[M,M^\tau]) \\ = \;& \|M-M^\tau\|^2 - \left(\|M\|^2 + \|M^\tau\|^2 - 2(M,M^\tau)\right) + \text{local martingale} \\ = \;& \text{local martingale.} \end{aligned}$$

By the first part of the corollary which ensures $[M,M^\tau]$ is of bounded variation, the left side is the difference of two increasing adapted processes and so by Corollary 32.3.5 again, the left side equals 0. Thus from the above,

$$\begin{aligned} [M-M^\tau] &= [M] + [M^\tau] - 2[M,M^\tau] = [M] + [M^\tau] - 2[M^\tau,M^\tau] \\ &= [M] + [M^\tau] - 2[M^\tau] = [M] - [M^\tau] \leq [M] \end{aligned}$$

Finally consider the claim that $[M,N]$ is bilinear. From the definition, letting M_1, M_2, N be H valued local martingales,

$$\begin{aligned} (aM_1 + bM_2, N)_H &= [aM_1 + bM_2, N] + \text{local martingale} \\ a(M_1,N) + b(M_2,N)_H &= a[M_1,N] + b[M_2,N] + \text{local martingale} \end{aligned}$$

Hence

$$[aM_1 + bM_2, N] - (a[M_1,N] + b[M_2,N]) = \text{local martingale.}$$

The left side can be written as the difference of two increasing functions thanks to $[M,N]$ of bounded variation and so by Lemma 32.2.1 it equals 0. $[M,N]$ is obviously symmetric from the definition. ■

32.5 The Burkholder Davis Gundy Inequality

Define

$$M^*(\omega) \equiv \sup\{\|M(t)(\omega)\| : t \in [0,T]\}.$$

The Burkholder Davis Gundy inequality is an amazing inequality which involves M^* and $[M](T)$.

Before presenting this, here is the good lambda inequality, Theorem 10.12.1 on Page 306 listed here for convenience.

Theorem 32.5.1 *Let $(\Omega, \mathscr{F}, \mu)$ be a finite measure space and let F be a continuous increasing function defined on $[0,\infty)$ such that $F(0) = 0$. Suppose also that for all $\alpha > 1$, there exists a constant C_α such that for all $x \in [0,\infty)$,*

$$F(\alpha x) \leq C_\alpha F(x).$$

Also suppose f, g are nonnegative measurable functions and there exists $\beta > 1, 0 < r \leq 1$, such that for all $\lambda > 0$ and $1 > \delta > 0$,

$$\mu([f > \beta\lambda] \cap [g \leq r\delta\lambda]) \leq \phi(\delta)\mu([f > \lambda]) \tag{32.13}$$

where $\lim_{\delta \to 0+} \phi(\delta) = 0$ and ϕ is increasing. Under these conditions, there exists a constant C depending only on β, ϕ, r such that

$$\int_\Omega F(f(\omega))\, d\mu(\omega) \leq C \int_\Omega F(g(\omega))\, d\mu(\omega).$$

The proof of the Burkholder Davis Gundy inequality also will depend on the hitting this before that theorem which is listed next for convenience.

Theorem 32.5.2 *Let $\{M(t)\}$ be a continuous real valued martingale adapted to the normal filtration $\mathcal{F}_t$ and let*

$$M^* \equiv \sup\{|M(t)| : t \geq 0\}$$

and $M(0) = 0$. Letting

$$\tau_x \equiv \inf\{t > 0 : M(t) = x\}$$

Then if $a < 0 < b$ the following inequalities hold.

$$(b-a)P([\tau_b \leq \tau_a]) \geq -aP([M^* > 0]) \geq (b-a)P([\tau_b < \tau_a])$$

and

$$(b-a)P([\tau_a < \tau_b]) \leq bP([M^* > 0]) \leq (b-a)P([\tau_a \leq \tau_b]).$$

In words, $P([\tau_b \leq \tau_a])$ is the probability that $M(t)$ hits b no later than when it hits a. (Note that if $\tau_a = \infty = \tau_b$ then you would have $[\tau_a = \tau_b]$.)

Then the Burkholder Davis Gundy inequality is as follows. Generalizations will be presented later.

Theorem 32.5.3 *Let $\{M(t)\}$ be a continuous H valued martingale which is uniformly bounded, $M(0) = 0$, where H is a separable Hilbert space and $t \in [0,T]$. Then if F is a function of the sort described in the good lambda inequality above, there are constants, C and c independent of such martingales M such that*

$$c\int_\Omega F\left(([M](T))^{1/2}\right)dP \leq \int_\Omega F(M^*)\,dP \leq C\int_\Omega F\left(([M](T))^{1/2}\right)dP$$

where

$$M^*(\omega) \equiv \sup\{\|M(t)(\omega)\| : t \in [0,T]\}.$$

Proof: Using Corollary 32.4.3, let

$$\begin{aligned} N(t) &\equiv \|M(t) - M^\tau(t)\|^2 - [M - M^\tau](t) \\ &= \|M(t) - M^\tau(t)\|^2 - [M](t) + [M]^\tau(t) \end{aligned}$$

where

$$\tau \equiv \inf\{t \in [0,T] : \|M(t)\| > \lambda\}$$

Thus N is a martingale and $N(0) = 0$. In fact $N(t) = 0$ as long as $t \leq \tau$. As usual $\inf(\emptyset) \equiv \infty$. Note

$$[\tau < \infty] = [M^* > \lambda] \supseteq [N^* > 0]$$

This is because to say $\tau < \infty$ is to say there exists $t < T$ such that $\|M(t)\| > \lambda$ which is the same as saying $M^* > \lambda$. Thus the first two sets are equal. Either $\tau < \infty$ or $\tau = \infty$. If $\tau = \infty$, then from the formula for $N(t)$ above, $N(t) = 0$ for all $t \in [0,T]$ and so it can't happen that $N^* > 0$. Thus $[\tau = \infty] \subseteq [N^* = 0]$ so $[N^* > 0] \subseteq [\tau < \infty]$.

Let $\beta > 2$ and let $\delta \in (0,1)$. Then $\beta - 1 > 1 > \delta > 0$. Consider the following which is set up to use the good lambda inequality.

$$S_r \equiv [M^* > \beta\lambda] \cap \left[([M](T))^{1/2} \leq r\delta\lambda \right]$$

where $0 < r < 1$.It is shown that S_r corresponds to hitting "this before that" and there is an estimate for this which involves $P([N^* > 0])$ which is bounded above by $P([M^* > \lambda])$ as discussed above. This will satisfy the hypotheses of the good lambda inequality.

Claim: For $\omega \in S_r$, $N(t)$ hits $\lambda^2\left(1-\delta^2\right)$.

Proof of claim: For $\omega \in S_r$, there exists a $t < T$ such that $\|M(t)\| > \beta\lambda$ and so using Corollary 32.4.3 and triangle inequality,

$$\begin{aligned} N(t) &\geq \left| \|M(t)\| - \|M^\tau(t)\| \right|^2 - [M - M^\tau](t) \geq |\beta\lambda - \lambda|^2 - [M](t) \\ &\geq (\beta-1)^2\lambda^2 - \delta^2\lambda^2 \end{aligned}$$

which shows that $N(t)$ hits $(\beta-1)^2\lambda^2 - \delta^2\lambda^2$ for $\omega \in S_r$. By the intermediate value theorem, it also hits $\lambda^2\left(1-\delta^2\right)$. This proves the claim.

Claim: $N(t)(\omega)$ never hits $-\delta^2\lambda^2$ for $\omega \in S_r$.

Proof of claim: Suppose t is the first time $N(t)$ reaches $-\delta^2\lambda^2$. Then $t > \tau$ because $N(t) = 0$ on $[0,\tau]$ and so

$$\begin{aligned} N(t) &= -\delta^2\lambda^2 \geq \left| \|M(t)\| - \lambda \right|^2 - [M](t) + [M^\tau](t) \\ &\geq -r^2\lambda^2\delta^2, \end{aligned}$$

a contradiction since $r < 1$. This proves the claim.

Therefore, for all $\omega \in S_r$, $N(t)(\omega)$ reaches $\lambda^2\left(1-\delta^2\right)$ before it reaches $-\delta^2\lambda^2$. It follows

$$P(S_r) \leq P\left(N(t) \text{ reaches } \lambda^2\left(1-\delta^2\right) \text{ before } -\delta^2\lambda^2\right)$$

and because of Theorem 31.6.3 this is no larger than

$$P([N^* > 0]) \frac{\delta^2\lambda^2}{\lambda^2\left(1-\delta^2\right) - \left(-\delta^2\lambda^2\right)} = P([N^* > 0])\,\delta^2 \leq \delta^2 P([M^* > \lambda]).$$

Thus

$$P\left([M^* > \beta\lambda] \cap \left[([M](T))^{1/2} \leq r\delta\lambda \right]\right) \leq P([M^* > \lambda])\,\delta^2$$

By the good lambda inequality,

$$\int_\Omega F(M^*)\,dP \leq C \int_\Omega F\left(([M](T))^{1/2} \right) dP$$

which is one half the inequality.

Now consider the other half. This time define the stopping time τ by

$$\tau \equiv \inf\left\{ t \in [0,T] : ([M](t))^{1/2} > \lambda \right\}$$

and let

$$S_r \equiv \left[([M](T))^{1/2} > \beta\lambda \right] \cap [2M^* \leq r\delta\lambda].$$

Then there exists $t < T$ such that $[M](t) > \beta^2\lambda^2$. This time, let

$$N(t) \equiv [M](t) - [M^\tau](t) - \|M(t) - M^\tau(t)\|^2$$

This is still a martingale since by Corollary 32.4.3

$$[M](t) - [M^\tau](t) = [M - M^\tau](t)$$

Claim: $N(t)(\omega)$ hits $\lambda^2\left(1-\delta^2\right)$ for some $t < T$ for $\omega \in S_r$.

Proof of claim: Fix such a $\omega \in S_r$. Let $t < T$ be such that $[M](t) > \beta^2\lambda^2$. Then, since $\beta > 2$, $t > \tau$ and so for that ω,

$$\begin{aligned} N(t) &> \beta^2\lambda^2 - \lambda^2 - \|M(t) - M(\tau)\|^2 \\ &\geq (\beta-1)^2\lambda^2 - (\|M(t)\| + \|M(\tau)\|)^2 \\ &\geq (\beta-1)^2\lambda^2 - r^2\delta^2\lambda^2 \geq \lambda^2 - \delta^2\lambda^2 \end{aligned}$$

By the intermediate value theorem, it hits $\lambda^2\left(1-\delta^2\right)$. The last inequality follows because it is assumed that $2M^* \leq r\delta\lambda$. This proves the claim.

Claim: $N(t)(\omega)$ never hits $-\delta^2\lambda^2$ for $\omega \in S_r$.

Proof of claim: By Corollary 32.4.3, if it did at t, then $t > \tau$ because $N(t) = 0$ for $t \leq \tau$, and so

$$\begin{aligned} 0 &\leq [M](t) - [M^\tau](t) = \|M(t) - M(\tau)\|^2 - \delta^2\lambda^2 \\ &\leq (\|M(t)\| + \|M(\tau)\|)^2 - \delta^2\lambda^2 \leq r^2\delta^2\lambda^2 - \delta^2\lambda^2 < 0, \end{aligned}$$

a contradiction. The last inequality follows from $2M^* \leq r\delta\lambda$ on S_r. This proves the claim.

It follows that for each $r \in (0,1)$,

$$P(S_r) \leq P\left(N(t) \text{ hits } \lambda^2\left(1-\delta^2\right) \text{ before } -\delta^2\lambda^2\right)$$

By Theorem 31.6.3 this is no larger than

$$P([N^* > 0])\frac{\delta^2\lambda^2}{\lambda^2\left(1-\delta^2\right)+\delta^2\lambda^2} = P([N^* > 0])\,\delta^2$$

$$\leq P([\tau < \infty])\,\delta^2 = P\left(\left[([M](T))^{1/2} > \lambda\right]\right)\delta^2$$

Now by the good lambda inequality, there is a constant k independent of M such that

$$\int_\Omega F\left(([M](T))^{1/2}\right)dP \leq k\int_\Omega F(2M^*)\,dP \leq kC_2\int_\Omega F(M^*)\,dP$$

by the assumptions about F. Therefore, combining this result with the first part,

$$\begin{aligned} (kC_2)^{-1}\int_\Omega F\left(([M](T))^{1/2}\right)dP &\leq \int_\Omega F(M^*)\,dP \\ &\leq C\int_\Omega F\left(([M](T))^{1/2}\right)dP \blacksquare \end{aligned}$$

Of course, everything holds for local martingales in place of martingales.

Theorem 32.5.4 *Let $\{M(t)\}$ be a continuous H valued local martingale, $M(0) = 0$, where H is a separable Hilbert space and $t \in [0,T]$. Then if F is a function of the sort described in the good lambda inequality, that is,*

$$F(0) = 0,\ F \text{ continuous, } F \text{ increasing,}$$

$$F(\alpha x) \leq c_\alpha F(x),$$

there are constants, C and c independent of such local martingales M such that

$$c\int_\Omega F\left([M](T)^{1/2}\right)dP \leq \int_\Omega F(M^*)\,dP \leq C\int_\Omega F\left([M](T)^{1/2}\right)dP$$

where

$$M^*(\omega) \equiv \sup\{\|M(t)(\omega)\| : t \in [0,T]\}.$$

Proof: Let $\{\tau_n\}$ be an increasing localizing sequence for M such that M^{τ_n} is uniformly bounded. Such a localizing sequence exists from Proposition 32.3.3. Then from Theorem 32.5.3 there exist constants c, C independent of τ_n such that

$$\begin{aligned} c\int_\Omega F\left([M^{\tau_n}](T)^{1/2}\right)dP &\leq \int_\Omega F\left((M^{\tau_n})^*\right)dP \\ &\leq C\int_\Omega F\left([M^{\tau_n}](T)^{1/2}\right)dP \end{aligned}$$

By Corollary 32.4.3, this implies

$$\begin{aligned} c\int_\Omega F\left(\left([M]^{\tau_n}\right)(T)^{1/2}\right)dP &\leq \int_\Omega F\left((M^{\tau_n})^*\right)dP \\ &\leq C\int_\Omega F\left(\left([M]^{\tau_n}\right)(T)^{1/2}\right)dP \end{aligned}$$

and now note that $\left([M]^{\tau_n}\right)(T)^{1/2}$ and $(M^{\tau_n})^*$ increase in n to $[M](T)^{1/2}$ and M^* respectively. Then the result follows from the monotone convergence theorem. ■

Here is a corollary [45].

Corollary 32.5.5 *Let $\{M(t)\}$ be a continuous H valued local martingale and let $\varepsilon, \delta \in (0,\infty)$. Then there is a constant C, independent of ε, δ such that*

$$P\left(\left[\overbrace{\sup_{t\in[0,T]}\|M(t)\|}^{M^*(T)} \geq \varepsilon\right]\right) \leq \frac{C}{\varepsilon}E\left([M]^{1/2}(T)\wedge\delta\right) + P\left([M]^{1/2}(T) > \delta\right)$$

Proof: Let the stopping time τ be defined by

$$\tau \equiv \inf\left\{t > 0 : [M]^{1/2}(t) > \delta\right\}$$

Then

$$P([M^* \geq \varepsilon]) = P([M^* \geq \varepsilon] \cap [\tau = \infty]) + P([M^* \geq \varepsilon] \cap [\tau < \infty])$$

On the set where $[\tau = \infty]$, $M^\tau = M$ and so $P([M^* \geq \varepsilon]) \leq$

$$\leq \frac{1}{\varepsilon}\int_\Omega (M^\tau)^*\,dP + P\left([M^* \geq \varepsilon] \cap \left[[M]^{1/2}(T) > \delta\right]\right)$$

By Theorem 32.5.4 and Corollary 32.4.3,

$$\begin{aligned} &\leq \frac{C}{\varepsilon}\int_\Omega [M^\tau]^{1/2}(T)\,dP + P\left([M^*\geq\varepsilon]\cap\left[[M]^{1/2}(T)>\delta\right]\right)\\ &= \frac{C}{\varepsilon}\int_\Omega \left([M]^\tau\right)^{1/2}(T)\,dP + P\left([M^*\geq\varepsilon]\cap\left[[M]^{1/2}(T)>\delta\right]\right)\\ &\leq \frac{C}{\varepsilon}\int_\Omega [M]^{1/2}(T)\wedge\delta\,dP + P\left([M^*\geq\varepsilon]\cap\left[[M]^{1/2}(T)>\delta\right]\right)\end{aligned}$$

$$\leq \frac{C}{\varepsilon}\int_\Omega [M]^{1/2}(T)\wedge\delta\,dP + P\left(\left[[M]^{1/2}(T)>\delta\right]\right) \blacksquare$$

The Burkholder Davis Gundy inequality along with the properties of the covariation implies the following amazing proposition.

Proposition 32.5.6 *The space $M_T^2(H)$ is a Hilbert space with respect to an equivalent norm. Here H is a separable Hilbert space.*

Proof: We already know from Proposition 31.7.2 that this space is a Banach space. It is only necessary to exhibit an equivalent norm which makes it a Hilbert space. However, you can let $F(\lambda)=\lambda^2$ in the Burkholder Davis Gundy theorem and obtain for $M\in M_T^2(H)$, the two norms

$$\left(\int_\Omega [M](T)\,dP\right)^{1/2} = \left(\int_\Omega [M,M](T)\,dP\right)^{1/2}$$

and

$$\left(\int_\Omega (M^*)^2\,dP\right)^{1/2}$$

are equivalent. The first comes from an inner product since from Corollary 32.4.3, $[\cdot,\cdot]$ is bilinear and symmetric and nonnegative. If $[M,M](T)=[M](T)=0$ in $L^1(\Omega)$, then from the Burkholder Davis Gundy inequality, $M^*=0$ in $L^2(\Omega)$ and so $M=0$. Hence

$$\int_\Omega [M,N](T)\,dP$$

is an inner product which yields the equivalent norm. $\blacksquare$

Later, the Wiener process will be discussed and the existence of such a process is proved. For now, the following example shows something about such processes.

Example 32.5.7 *An example of a real martingale is the Wiener process $W(t)$. It has the property that whenever $t_1<t_2<\cdots<t_n$, the increments $\{W(t_i)-W(t_{i-1})\}$ are independent and whenever $s<t$, $W(t)-W(s)$ is normally distributed with mean 0 and variance $(t-s)$. For the Wiener process, we let*

$$\mathscr{F}_t \equiv \cap_{u>t}\overline{\sigma(W(s)-W(r):r<s\leq u)}$$

and it is with respect to this normal filtration that W is a continuous martingale. What is the quadratic variation of such a process?

The quadratic variation of the Wiener process is just t. This is because if $A \in \mathscr{F}_s, s < t$,

$$E\left(\mathscr{X}_A\left(|W(t)|^2 - t\right)\right) =$$

$$E\left(\mathscr{X}_A\left(|W(t) - W(s)|^2 + |W(s)|^2 + 2(W(s), W(t) - W(s)) - (t - s + s)\right)\right)$$

Now

$$E(\mathscr{X}_A(2(W(s), W(t) - W(s)))) = P(A) E(2W(s)) E(W(t) - W(s)) = 0$$

by the independence of the increments. Thus the above reduces to

$$E\left(\mathscr{X}_A\left(|W(t) - W(s)|^2 + |W(s)|^2 - (t - s + s)\right)\right)$$

$$\begin{aligned} &= E\left(\mathscr{X}_A\left(|W(t) - W(s)|^2 - (t - s)\right)\right) + E\left(\mathscr{X}_A\left(|W(s)|^2 - s\right)\right) \\ &= P(A) E\left(|W(t) - W(s)|^2 - (t - s)\right) + E\left(\mathscr{X}_A\left(|W(s)|^2 - s\right)\right) \\ &= E\left(\mathscr{X}_A\left(|W(s)|^2 - s\right)\right) \end{aligned}$$

and so $E\left(|W(t)|^2 - t | \mathscr{F}_s\right) = |W(s)|^2 - s$ showing that $t \to |W(t)|^2 - t$ is a martingale. Hence, by uniqueness, $[W](t) = t$.

32.6 Approximation With Step Functions

There is a really nice result about approximating a function $f \in L^p([0,T], E)$ with step functions. In this we deal with a specific representative of the equivalence class for $f \in L^p([0,T], E)$.

Lemma 32.6.1 *Let $f \in L^2([0,T]; E)$ for E a Banach space. For simplicity let f be Borel measurable. Then there exists a sequence of nested partitions, $\mathscr{P}_k \subseteq \mathscr{P}_{k+1}$,*

$$\mathscr{P}_k \equiv \left\{t_0^k, \cdots, t_{m_k}^k\right\}$$

such that the step functions given by

$$\begin{aligned} f_k^r(t) &\equiv \sum_{j=1}^{m_k} f\left(t_j^k\right) \mathscr{X}_{[t_{j-1}^k, t_j^k)}(t) \\ f_k^l(t) &\equiv \sum_{j=1}^{m_k} f\left(t_{j-1}^k\right) \mathscr{X}_{[t_{j-1}^k, t_j^k)}(t) \end{aligned}$$

both converge to f in $L^2([0,T]; E)$ as $k \to \infty$ and

$$\lim_{k \to \infty} \max\left\{\left|t_j^k - t_{j+1}^k\right| : j \in \{0, \cdots, m_k\}\right\} = 0.$$

The mesh points $\left\{t_j^k\right\}_{j=0}^{m_k}$ can be chosen to miss a given set of measure zero N if N does not contan either 0 or T.

Note that it would make no difference in terms of the conclusion of this lemma if you defined

$$f_k^l(t) \equiv \sum_{j=1}^{m_k} f\left(t_{j-1}^k\right) \mathcal{X}_{(t_{j-1}^k, t_j^k]}(t)$$

because the modified function equals the one given above off a countable subset of $[0,T]$, the union of the mesh points. One could change f_k^r similarly with no change in the conclusion.

Proof: Let f be 0 off $(0,T)$. Thus we will let it be defined on all of $\mathbb{R}$. Let $\gamma_n(t) \equiv k/2^n, \delta_n(t) \equiv (k+1)/2^n$, where $t \in (k/2^n, (k+1)/2^n]$, and $2^{-n} < \delta$. Thus $\gamma_n(t)$ is the closest $k2^{-n}, k \in \mathbb{Z}$ which is smaller than or equal to t while $\delta_n(t)$ is the closest $k2^{-n}$ larger than or equal to t. Let $g \in C_c((0,T),E)$ so the support of g is in $[\delta, T-\delta]$ for some $\delta > 0$ such that $\int_{[\delta,T-\delta]^C} \|f\|^p\, dt < \varepsilon$ and also

$$\int_0^T \|f-g\|^p\, dt = \int_{\mathbb{R}} \|f-g\|^p\, dt < \varepsilon.$$

Then

$$\int_0^T \int_0^T \|f(\gamma_n(u-s)+s) - g(\gamma_n(u-s)+s)\|_E^p\, dsdu$$

$$\begin{aligned} &= \int_0^T \int_0^T \|f(\gamma_n(u-s)+s) - g(\gamma_n(u-s)+s)\|_E^p\, duds \\ &= \int_0^T \int_{-s}^{T-s} \|f(\gamma_n(t)+s) - g(\gamma_n(t)+s)\|_E^p\, dtds \end{aligned}$$

$$\begin{aligned} &\leq \int_0^T \int_{-2T}^{2T} \|f(\gamma_n(t)+s) - g(\gamma_n(t)+s)\|_E^p\, dtds \\ &= \int_{-2T}^{2T} \int_0^T \|f(\gamma_n(t)+s) - g(\gamma_n(t)+s)\|_E^p\, dsdt \\ &\leq \int_{-2T}^{2T} \int_{\mathbb{R}} \|f(\gamma_n(t)+s) - g(\gamma_n(t)+s)\|_E^p\, dsdt < 5T\varepsilon \end{aligned}$$

No effort is made to get the best possible estimate in the above. Then

$$\int_0^T \int_0^T \|f(\gamma_n(u-s)+s) - f(u) - (g(\gamma_n(u-s)+s) - g(u))\|_E^p\, duds$$

$$\leq 2^{p-1} \int_0^T \int_0^T \left(\begin{array}{c} \|f(\gamma_n(u-s)+s) - g(\gamma_n(u-s)+s)\|_E^p \\ + \|f(u) - g(u)\|_E^p \end{array} \right) duds$$

$$\leq 2^{p-1}\varepsilon 5T + 2^{p-1}\varepsilon T = 2^{p-1}6\varepsilon T \tag{32.14}$$

It follows that if n is chosen large enough, then

$$\int_0^T \int_0^T \|g(\gamma_n(u-s)+s) - g(u)\|_E^p\, duds < \varepsilon T$$

from uniform continuity of g. Therefore, from 32.14

$$\begin{aligned}&\int_0^T\int_0^T \|f(\gamma_n(u-s)+s)-f(u)\|^p\,duds\\ \leq\ & 2^{p-1}\int_0^T\int_0^T \|g(\gamma_n(u-s)+s)-g(u)\|^p\,duds\end{aligned}$$

$$+2^{p-1}\int_0^T\int_0^T \|f(\gamma_n(u-s)+s)-f(u)-(g(\gamma_n(u-s)+s)-g(u))\|_E^p\,duds$$

$$\leq 2^{p-1}\varepsilon T+2^{p-1}\left(2^{p-1}6\varepsilon T\right) \tag{32.15}$$

Now $\gamma_n(u-s)+s\geq 0$ unless $u<\delta$. So consider

$$\int_0^T\int_0^T \left\|f\left((\gamma_n(u-s)+s)^+\right)-f(u)\right\|^p\,duds\leq$$

$$\int_0^T\int_{2^{-n}}^T \|f(\gamma_n(u-s)+s)-f(u)\|^p\,duds+\int_0^T\int_0^{2^{-n}} \|f(u)\|^p\,duds$$

$$\leq 2^{p-1}\varepsilon T+2^{p-1}\left(2^{p-1}6\varepsilon T\right)+\varepsilon\equiv\eta$$

Thus, since ε is arbitrary, so is η.

The function $u\to(\gamma_n(u-s)+s)^+$ has jumps $0=t_0,t_1,t_2,....t_{m_n-1},t_{m_n}=T$ where these are listed in increasing order. The possible values of these t_i are $k2^{-n}+s$ for some $k\in\mathbb{Z}$. They are equally spaced being 2^{-n} apart except for the first two and the last two which are no more than 2^{-n}. One can slide this list of partition points around according to the choice of $s\in[0,T]$. Now suppose you have a set of measure zero N. Pick $s\in[0,T]$ such that none of the t_i are in N and

$$\int_0^T \left\|f\left((\gamma_n(u-s)+s)^+\right)-f(u)\right\|_E^p\,du<2\eta$$

Now let $f_n(u)\equiv\sum_{k=1}^{m_n} f(t_{k-1})\,\mathcal{X}_{[t_{k-1},t_k)}(u)$. This is a step function of the desired sort. Then

$$\int_0^T \|f_n(u)-f(u)\|_E^p\,du=\sum_{k=1}^{m_n}\int_{t_{k-1}}^{t_k} \|f(t_{k-1})-f(u)\|_E^p\,du$$

$$\begin{aligned}&=\sum_{k=1}^{m_n}\int_{t_{k-1}}^{t_k} \left\|f\left((\gamma_n(u-s)+s)^+\right)-f(u)\right\|_E^p\,du\\ &=\int_0^T \left\|f\left((\gamma_n(u-s)+s)^+\right)-f(u)\right\|_E^p\,du<2\eta\end{aligned}$$

Picking a sequence of these step functions f_j corresponding to $\eta=2^{-j}$, one obtains the desired sequence in which values of f are assigned at the left end point of the interval. Making n still larger in the above argument and using the same argument with the right end points, one can also obtain a similar step function in which the values of f are given at the right end point which also converges to f in $L^2([0,T],E)$. ■

Chapter 33

Quadratic Variation and Stochastic Integration

Let $\mathscr{F}_t$ be a normal filtration and let $\{M(t)\}$ be a continuous local martingale adapted to $\mathscr{F}_t$ having values in U a separable real Hilbert space.

Definition 33.0.1 *Let $\mathscr{F}_t$ be a normal filtration and let*

$$f(t) \equiv \sum_{k=0}^{m_n-1} f_k \mathscr{X}_{(t_k,t_{k+1}]}(t)$$

where $\{t_k\}_{k=0}^{m_n}$ is a partition of $[0,T]$ and each f_k is $\mathscr{F}_{t_k}$ measurable, $f_k M^* \in L^2(\Omega)$ where

$$M^*(\omega) \equiv \sup_{t\in[0,T]} \|M(t)(\omega)\|$$

Such a function is called an elementary function. Also let $\{M(t)\}$ be a continuous local martingale adapted to $\mathscr{F}_t$ which has values in a separable real Hilbert space U such that $M(0) = 0$. For such an elementary real valued function, define

$$\int_0^t f dM \equiv \sum_{k=0}^{m_n-1} f_k \left(M(t\wedge t_{k+1}) - M(t\wedge t_k)\right). \tag{33.1}$$

Since the $t \to \mathscr{F}_t$ is increasing, this definition is well defined. Also the set of elementary functions is a vector space.

Then with this definition, here is a wonderful lemma.

Lemma 33.0.2 *For f an elementary function as above, $\left\{\int_0^t f dM\right\}$ is a continuous local martingale and*

$$E\left(\left\|\int_0^t f dM\right\|_U^2\right) = \int_\Omega \int_0^t f(s)^2 d[M](s) dP. \tag{33.2}$$

If N is another continuous local martingale adapted to $\mathscr{F}_t$ and both f, g are elementary functions such that for each k,

$$f_k M^*, g_k N^* \in L^2(\Omega),$$

then

$$E\left(\left(\int_0^t f dM, \int_0^t g dN\right)_U\right) = \int_\Omega \int_0^t f g d[M,N] \tag{33.3}$$

and both sides make sense.

Proof: Let $\{\tau_l\}$ be a localizing sequence for M such that M^{τ_l} is a bounded martingale. Then from the definition, for each ω

$$\int_0^t f dM = \lim_{l\to\infty} \int_0^t f dM^{\tau_l} = \lim_{l\to\infty} \left(\int_0^t f dM\right)^{\tau_l}$$

and it is clear that $\left\{\int_0^t f dM^{\tau_l}\right\}$ is a martingale because it is just the sum of some martingales. Thus $\{\tau_l\}$ is a localizing sequence for $\int_0^t f dM$. It is also clear $\int_0^t f dM$ is continuous

because it is a finite sum of continuous random variables. In the argument to get 33.2, I will write M rather than M^{τ_l}. Then it is understood that you can let $l \to \infty$ to obtain the desired formula for a local martingale. Thus, in what follows, M is a bounded martingale.

$$\begin{aligned} & E\left(\left\|\int_0^t f dM\right\|_U^2\right) \\ = \; & E\left(\sum_{k=0}^{m_n-1} f_k\left(M\left(t\wedge t_{k+1}\right)-M\left(t\wedge t_k\right)\right), \sum_{j=0}^{m_n-1} f_j\left(M\left(t\wedge t_{j+1}\right)-M\left(t\wedge t_j\right)\right)\right) \end{aligned}$$

Let $M\left(t\wedge t_{k+1}\right)-M\left(t\wedge t_k\right)=\Delta M_k$. Thus ΔM_k is $\mathscr{F}_{t_{k+1}}$ measurable. Consider a mixed term in the above in which $j<k$

$$\begin{aligned} E\left(f_k\Delta M_k, f_j\Delta M_j\right) &= E\left(E\left(f_k\Delta M_k, f_j\Delta M_j|\mathscr{F}_{t_k}\right)\right) \\ &= E\left(f_k, f_j\Delta M_j E\left(\Delta M_k|\mathscr{F}_{t_k}\right)\right)=0 \end{aligned}$$

because $E\left(M\left(t\wedge t_{k+1}\right)-M\left(t\wedge t_k\right)|\mathscr{F}_{t_k}\right)=M\left(t\wedge t_{k+1}\wedge t_k\right)-M\left(t\wedge t_k\right)$. Thus

$$\begin{aligned} E\left(\left\|\int_0^t f dM\right\|_U^2\right) &= \sum_{k=0}^{m_n-1} E\left(f_k\Delta M_k, f_k\Delta M_k\right) \\ &= \sum_{k=0}^{m_n-1} E\left(E\left(\left(f_k\Delta M_k, f_k\Delta M_k\right)|\mathscr{F}_{t_k}\right)\right) \qquad (33.4) \\ &= \sum_{k=0}^{m_n-1} E\left(f_k^2 E\left(\left(\Delta M_k, \Delta M_k\right)|\mathscr{F}_{t_k}\right)\right) \qquad (33.5) \end{aligned}$$

now

$$\begin{aligned} E\left(\left(M\left(t\wedge t_{k+1}\right), M\left(t\wedge t_k\right)\right)|\mathscr{F}_{t_k}\right) &= \left(E\left(M\left(t\wedge t_{k+1}\right)|\mathscr{F}_{t_k}\right), M\left(t\wedge t_k\right)\right) \\ &= \left\|M\left(t\wedge t_k\right)\right\|^2 \end{aligned}$$

and so

$$E\left(\left(\Delta M_k, \Delta M_k\right)|\mathscr{F}_{t_k}\right)=E\left(\begin{pmatrix} M\left(t\wedge t_{k+1}\right)-M\left(t\wedge t_k\right), \\ M\left(t\wedge t_{k+1}\right)-M\left(t\wedge t_k\right) \end{pmatrix}|\mathscr{F}_{t_k}\right)=$$

$$\begin{aligned} & E\left(\left\|M\left(t\wedge t_{k+1}\right)\right\|^2|\mathscr{F}_{t_k}\right)+E\left(\left\|M\left(t\wedge t_k\right)\right\|^2|\mathscr{F}_{t_k}\right) \\ & -2E\left(\left(M\left(t\wedge t_{k+1}\right), M\left(t\wedge t_k\right)\right)|\mathscr{F}_{t_k}\right) \\ = \; & E\left(\left\|M\left(t\wedge t_{k+1}\right)\right\|^2|\mathscr{F}_{t_k}\right)-\left\|M\left(t\wedge t_k\right)\right\|^2 \end{aligned}$$

Therefore, the right side of 33.5 is

$$\sum_{k=0}^{m_n-1} E\left(f_k^2\left\|M\left(t\wedge t_{k+1}\right)\right\|^2\right)-E\left(f_k^2\left\|M\left(t\wedge t_k\right)\right\|^2\right).$$

Now recall that $\|M(t)\|^2=[M](t)+N(t)$ where $N(t)$ is a martingale. It then reduces to

$$\sum_{k=0}^{m_n-1} E\left(f_k^2\left(\left[M\left(t\wedge t_{k+1}\right)\right]+N\left(t\wedge t_{k+1}\right)\right)\right)-E\left(f_k^2\left([M]\left(t\wedge t_k\right)+N\left(t\wedge t_k\right)\right)\right)$$

$$= \int_\Omega \int_0^t f(s)^2 \, d[M] \, dP$$

since the martingales integrate to 0. This proves the first formula. If $f(s) \in \mathscr{L}(U,H)$, you could probably modify this argument. In this case, you couldn't factor out of the inner product because it would no longer be a scalar. However, if $f(s) \in \mathscr{L}_2(U,H)$ you maybe could do something. This is really a version of the Ito isometry discussed later.

Next is a similar argument for the case of two different elementary functions. There is no loss of generality in assuming the mesh points are the same for the two elementary functions because if not, one can simply add in points to make this happen. It suffices to consider 33.3 because the other formula is a special case. To begin with, let $\{\tau_l\}$ be a localizing sequence which makes both M^{τ_l} and N^{τ_l} into bounded martingales. Consider the stopped process.

$$E\left(\left(\int_0^t f dM^{\tau_l}, \int_0^t g dN^{\tau_l}\right)_U\right)$$

$$= E\left(\left(\sum_{k=0}^{m_n-1} f_k\left(M^{\tau_l}(t\wedge t_{k+1}) - M^{\tau_l}(t\wedge t_k)\right), \right.\right.$$
$$\left.\left.\sum_{k=0}^{m_n-1} g_k\left(N^{\tau_l}(t\wedge t_{k+1}) - N^{\tau_l}(t\wedge t_k)\right)\right)\right)$$

To save on notation, write $M^{\tau_l}(t\wedge t_{k+1}) - M^{\tau_l}(t\wedge t_k) \equiv \Delta M_k(t)$, similar for ΔN_k. Thus

$$\Delta M_k = M^{\tau_l \wedge t_{k+1}} - M^{\tau_l \wedge t_k},$$

similar for ΔN_k. Then the above equals

$$E\left(\sum_{k=0}^{m_n-1}\left(f_k\Delta M_k, \sum_{k=0}^{m_n-1} g_k \Delta N_k\right)\right) = E\left(\sum_{k,j} f_k g_j (\Delta M_k, \Delta N_j)\right)$$

Now consider one of the mixed terms with $j<k$.

$$E((f_k\Delta M_k, g_j\Delta N_j)) = E\left(E\left((f_k\Delta M_k, g_j\Delta N_j)\,|\mathscr{F}_{t_k}\right)\right)$$
$$= E\left(g_j\Delta N_j, f_k E\left(\Delta M_k|\mathscr{F}_{t_k}\right)\right) = 0$$

since $E\left(\Delta M_k|\mathscr{F}_{t_k}\right) = E\left(\left(M^{\tau_l}(t\wedge t_{k+1}) - M^{\tau_l}(t\wedge t_k)\right)|\mathscr{F}_{t_k}\right) = 0$ by the Doob optional sampling theorem. Thus

$$E\left(\left(\int_0^t f dM^{\tau_l}, \int_0^t g dN^{\tau_l}\right)_U\right) = \tag{33.6}$$

$$= \sum_{k=0}^{m_n-1} E(f_k g_k(\Delta M_k, \Delta N_k)) = \sum_{k=0}^{m_n-1} E(f_k g_k([\Delta M_k, \Delta N_k] + \mathscr{N}_k)) \tag{33.7}$$

where $\mathscr{N}_k$ is a martingale such that $\mathscr{N}_k(t) = 0$ for all $t \le t_k$. This is because the martingale $(N^{\tau_l})^{t_{k+1}} - (N^{\tau_l})^{t_k} = \Delta N_k$ equals 0 for such t; and so $E(\mathscr{N}_k(t)) = 0$. Thus $f_k g_k \mathscr{N}_k$ is a martingale which equals zero when $t=0$. Therefore, its expectation also equals 0. Consequently the above reduces to

$$\sum_{k=0}^{m_n-1} E(f_k g_k [\Delta M_k, \Delta N_k]).$$

At this point, recall the definition of the covariation. The above equals

$$\frac{1}{4}\sum_{k=0}^{m_n-1} E\left(f_k g_k\left(\left[\Delta M_k+\Delta N_k\right]-\left[\Delta M_k-\Delta N_k\right]\right)\right)$$

Rewriting this yields

$$\begin{aligned}= \quad & \frac{1}{4}\sum_{k=0}^{m_n-1} E\left(f_k g_k\left(\left[\left(M^{\tau_l}\right)^{t_{k+1}}+\left(N^{\tau_l}\right)^{t_{k+1}}-\left(\left(M^{\tau_l}\right)^{t_k}+\left(N^{\tau_l}\right)^{t_k}\right)\right]\right.\right.\\ & \left.\left.-\left[\left(M^{\tau_l}\right)^{t_{k+1}}-\left(N^{\tau_l}\right)^{t_{k+1}}-\left(\left(M^{\tau_l}\right)^{t_k}-\left(N^{\tau_l}\right)^{t_k}\right)\right]\right)\right)\end{aligned}$$

To save on notation, denote

$$\begin{aligned}\left(M^{\tau_l}\right)^{t_{k+1}}+\left(N^{\tau_l}\right)^{t_{k+1}}-\left(\left(M^{\tau_l}\right)^{t_k}+\left(N^{\tau_l}\right)^{t_k}\right) &\equiv \Delta_k\left(M^{\tau_l}+N^{\tau_l}\right)\\ \left(M^{\tau_l}\right)^{t_{k+1}}-\left(N^{\tau_l}\right)^{t_{k+1}}-\left(\left(M^{\tau_l}\right)^{t_k}-\left(N^{\tau_l}\right)^{t_k}\right) &\equiv \Delta_k\left(M^{\tau_l}-N^{\tau_l}\right)\end{aligned}$$

Thus the above equals

$$\frac{1}{4}\sum_{k=0}^{m_n-1} E\left(f_k g_k\left(\left[\Delta_k\left(M^{\tau_l}+N^{\tau_l}\right)\right]-\left[\Delta_k\left(M^{\tau_l}-N^{\tau_l}\right)\right]\right)\right)$$

Now from Corollary 32.4.3,

$$=\frac{1}{4}\sum_{k=0}^{m_n-1} E\left(f_k g_k\left(\left[\Delta_k\left(M+N\right)\right]^{\tau_l}-\left[\Delta_k\left(M-N\right)\right]^{\tau_l}\right)\right)$$

Letting $l\to\infty$, this reduces to

$$\begin{aligned}= \quad & \frac{1}{4}\sum_{k=0}^{m_n-1} E\left(f_k g_k\left(\left[\Delta_k\left(M+N\right)\right]-\left[\Delta_k\left(M-N\right)\right]\right)\right)\\ = \quad & \frac{1}{4}\left(\int_\Omega\int_0^t fg\left(d\left[M+N\right]-d\left[M-N\right]\right)\right)\\ = \quad & \int_\Omega\int_0^t fg\,d\left[M,N\right]\end{aligned}$$

Now consider the left side of 33.7.

$$E\left(\left(\int_0^t f dM^{\tau_l},\int_0^t g dN^{\tau_l}\right)_U\right)$$

$$\begin{aligned}\equiv \quad & \int_\Omega\sum_{k,j} f_k g_j\left(\left(M^{\tau_l}\left(t\wedge t_{k+1}\right)-M^{\tau_l}\left(t\wedge t_k\right)\right),\right.\\ & \left.\left(N^{\tau_l}\left(t\wedge t_{j+1}\right)-N^{\tau_l}\left(t\wedge t_j\right)\right)\right)dP\end{aligned}$$

Then for each ω, the integrand converges as $l\to\infty$ to

$$\sum_{k,j} f_k g_j\left(\left(M\left(t\wedge t_{k+1}\right)-M\left(t\wedge t_k\right)\right),\left(N\left(t\wedge t_{j+1}\right)-N\left(t\wedge t_j\right)\right)\right)$$

But also you can do a sloppy estimate which will allow the use of the dominated convergence theorem.

$$\left\| \sum_{k,j} f_k g_j \left(M^{\tau_l}(t \wedge t_{k+1}) - M^{\tau_l}(t \wedge t_k) \right), \left(N^{\tau_l}\left(t \wedge t_{j+1}\right) - N^{\tau_l}(t \wedge t_j) \right) \right\|$$

$$\leq \sum_{k,j} |f_k| \left| g_j \right| 4M^* N^* \in L^1(\Omega)$$

by assumption. Thus the left side of 33.7 converges as $l \to \infty$ to

$$\int_\Omega \sum_{k,j} f_k g_j \left(\left(M(t \wedge t_{k+1}) - M(t \wedge t_k) \right), \left(N\left(t \wedge t_{j+1}\right) - N(t \wedge t_j) \right) \right) dP$$

$$= \int_\Omega \left(\int_0^t f dM, \int_0^t g dN \right)_U dP \blacksquare$$

Note for each ω, the inside integral in 33.2 is just a Stieltjes integral taken with respect to the increasing integrating function $[M]$.

Of course, with this estimate it is obvious how to extend the integral to a larger class of functions.

Definition 33.0.3 *Let $\nu(\omega)$ denote the Radon measure representing the functional*

$$\Lambda(\omega)(g) \equiv \int_0^T g d[M](t)(\omega)$$

($t \to [M](t)(\omega)$ is a continuous increasing function and $\nu(\omega)$ is the measure representing the Stieltjes integral, one for each ω.) Then let $\mathscr{G}_M$ denote functions $f(s,\omega)$ which are the limit of such elementary functions in the space $L^2\left(\Omega; L^2([0,T], \nu(\cdot))\right)$, the norm of such functions being

$$\|f\|^2_{\mathscr{G}} \equiv \int_\Omega \int_0^T f(s)^2 d[M](s) dP$$

For $f \in \mathscr{G}$ just defined,

$$\int_0^t f dM \equiv \lim_{n\to\infty} \int_0^t f_n dM$$

where $\{f_n\}$ is a sequence of elementary functions converging to f in

$$L^2\left(\Omega; L^2([0,T], \nu(\cdot))\right).$$

Now here is an interesting lemma.

Lemma 33.0.4 *Let M, N be continuous local martingales, $M(0) = N(0) = 0$ having values in a separable Hilbert space, U. Then*

$$[M+N]^{1/2} \leq \left([M]^{1/2} + [N]^{1/2} \right) \tag{33.8}$$

$$[M+N] \leq 2([M] + [N]) \tag{33.9}$$

Also, let ν_{M+N} denote the measure obtained from the increasing function $[M+N]$ and ν_N, ν_M be defined similarly,

$$\nu_{M+N} \leq 2(\nu_M + \nu_N) \tag{33.10}$$

on all Borel sets.

Proof: Since $(M,N) \to [M,N]$ is bilinear and satisfies

$$\begin{aligned} [M,N] &= [N,M] \\ [aM+bM_1,N] &= a[M,N]+b[M_1,N] \\ [M,M] &\geq 0 \end{aligned}$$

which follows from Corollary 32.4.3, the usual Cauchy Schwarz inequality holds and so

$$|[M,N]| \leq [M]^{1/2}[N]^{1/2}$$

Thus

$$\begin{aligned} [M+N] &\equiv [M+N,M+N] = [M,M]+[N,N]+2[M,N] \\ &\leq [M]+[N]+2[M]^{1/2}[N]^{1/2} = \left([M]^{1/2}+[N]^{1/2}\right)^2 \end{aligned}$$

This proves 33.8. Now square both sides. Then the right side is no larger than

$$2([M]+[N])$$

and this shows 33.9.

Now consider the claim about the measures. It was just shown that

$$[(M+N)-(M+N)^s] \leq 2([M-M^s]+[N-N^s])$$

and from Corollary 32.4.3 this implies that for $t > s$

$$\begin{aligned} & [M+N](t)-[M+N](s\wedge t) \\ = \; & [M+N](t)-[M+N]^s(t) \\ = \; & [M+N-(M^s+N^s)](t) \\ = \; & [M-M^s+(N-N^s)](t) \\ \leq \; & 2[M-M^s](t)+2[N-N^s](t) \\ \leq \; & 2([M](t)-[M](s))+2([N](t)-[N](s)) \end{aligned}$$

Thus

$$\nu_{M+N}([s,t]) \leq 2(\nu_M([s,t])+\nu_N([s,t]))$$

By regularity of the measures, this continues to hold with any Borel set F in place of $[s,t]$. ■

Theorem 33.0.5 *The integral is well defined and has a continuous version which is a local martingale. Furthermore it satisfies the Ito isometry,*

$$E\left(\left\|\int_0^t f\,dM\right\|_U^2\right) = \int_\Omega \int_0^t f(s)^2\,d[M](s)\,dP \tag{33.11}$$

Let the norm on $\mathcal{G}_N \cap \mathcal{G}_M$ be the maximum of the norms on $\mathcal{G}_N$ and $\mathcal{G}_M$ and denote by $\mathcal{E}_N$ and $\mathcal{E}_M$ the elementary functions corresponding to the martingales N and M respectively. Define $\mathcal{G}_{NM}$ as the closure in $\mathcal{G}_N \cap \mathcal{G}_M$ of $\mathcal{E}_N \cap \mathcal{E}_M$. Then for $f,g \in \mathcal{G}_{NM}$,

$$E\left(\left(\int_0^t f\,dM, \int_0^t g\,dN\right)\right) = \int_\Omega \int_0^t fg\,d[M,N] \tag{33.12}$$

Proof: It is clear the definition is well defined because if $\{f_n\}$ and $\{g_n\}$ are two sequences of elementary functions converging to f in $L^2\left(\Omega;L^2\left([0,T],\nu\right)\right)$ and if $\int_0^{1t} f dM$ is the integral which comes from $\{g_n\}$,

$$\begin{aligned}&\int_\Omega \left\| \int_0^{1t} f dM - \int_0^t f dM \right\|^2 dP \\ = \;& \lim_{n\to\infty} \int_\Omega \left\| \int_0^t g_n dM - \int_0^t f_n dM \right\|^2 dP \\ \leq \;& \lim_{n\to\infty} \int_\Omega \int_0^T \|g_n - f_n\|^2 d\nu dP = 0.\end{aligned}$$

Consider the claim the integral has a continuous version. Recall Theorem 31.4.3, part of which is listed here for convenience.

Theorem 33.0.6 *Let $\{X(t)\}$ be a right continuous nonnegative submartingale adapted to the normal filtration $\mathscr{F}_t$ for $t \in [0,T]$. Let $p \geq 1$. Define*

$$X^*(t) \equiv \sup\{X(s) : 0 < s < t\},\ X^*(0) \equiv 0.$$

Then for $\lambda > 0$

$$P([X^*(T) > \lambda]) \leq \frac{1}{\lambda^p} \int_\Omega X(T)^p dP \tag{33.13}$$

Let $\{f_n\}$ be a sequence of elementary functions converging to f in

$$L^2\left(\Omega;L^2\left([0,T],\nu(\cdot)\right)\right).$$

Then letting

$$X^{\tau_l}_{n,m}(t) = \left\| \int_0^t (f_n - f_m) dM^{\tau_l} \right\|_U,$$

$$X_{n,m}(t) = \left\| \int_0^t (f_n - f_m) dM \right\|_U = \left\| \int_0^t f_n dM - \int_0^t f_m dM \right\|_U$$

It follows $X^{\tau_l}_{n,m}$ is a continuous nonnegative submartingale and from Theorem 31.4.3 just listed,

$$P\left(\left[X^{\tau_l *}_{n,m}(T) > \lambda\right]\right) \leq \frac{1}{\lambda^2} \int_\Omega X^{\tau_l}_{n,m}(T)^2 dP$$

$$\begin{aligned}\leq \;& \frac{1}{\lambda^2} \int_\Omega \int_0^T |f_n - f_m|^2 d[M^{\tau_l}] dP \\ \leq \;& \frac{1}{\lambda^2} \int_\Omega \int_0^T |f_n - f_m|^2 d[M] dP\end{aligned}$$

Letting $l \to \infty$,

$$P\left(\left[X^*_{n,m}(T) > \lambda\right]\right) \leq \frac{1}{\lambda^2} \int_\Omega \int_0^T |f_n - f_m|^2 d[M] dP$$

Therefore, there exists a subsequence, still denoted by $\{f_n\}$ such that

$$P\left(\left[X^*_{n,n+1}(T) > 2^{-n}\right]\right) < 2^{-n}$$

Then by the Borel Cantelli lemma, the ω in infinitely many of the sets

$$\left[X^*_{n,n+1}(T) > 2^{-n}\right]$$

has measure 0. Denoting this exceptional set as N, it follows that for $\omega \notin N$, there exists $n(\omega)$ such that for $n > n(\omega)$,

$$\sup_{t\in[0,T]} \left\| \int_0^t f_n dM - \int_0^t f_{n+1} dM \right\| \leq 2^{-n}$$

and this implies uniform convergence of $\left\{\int_0^t f_n dM\right\}$. Letting

$$G(t) = \lim_{n\to\infty} \int_0^t f_n dM,$$

for $\omega \notin N$ and $G(t) = 0$ for $\omega \in N$, it follows that for each t, the continuous adapted process $G(t)$ equals $\int_0^t f dM$ a.e. Thus $\left\{\int_0^t f dM\right\}$ has a continuous version.

It suffices to verify 33.12. Let $\{f_n\}$ and $\{g_n\}$ be sequences of elementary functions converging to f and g in $\mathcal{G}_M \cap \mathcal{G}_N$. By Lemma 33.0.2,

$$E\left(\left(\int_0^t f_n dM, \int_0^t g_n dN\right)_U\right) = \int_\Omega \int_0^t f_n g_n d[M,N]$$

Then by the Holder inequality and the above definition,

$$\lim_{n\to\infty} E\left(\left(\int_0^t f_n dM, \int_0^t g_n dN\right)_U\right) = E\left(\left(\int_0^t f dM, \int_0^t g dN\right)_U\right)$$

Consider the right side which equals

$$\frac{1}{4}\int_\Omega \int_0^t f_n g_n d[M+N]\, dP - \frac{1}{4}\int_\Omega \int_0^t f_n g_n d[M-N]\, dP$$

Now from Lemma 33.0.4,

$$\begin{aligned} & \left| \int_\Omega \int_0^t f_n g_n d[M+N]\, dP - \int_\Omega \int_0^t fg\, d[M+N]\, dP \right| \\ = & \left| \int_\Omega \int_0^t f_n g_n d\nu_{M+N} dP - \int_\Omega \int_0^t fg\, d\nu_{M+N} dP \right| \end{aligned}$$

$$\leq 2\left(\int_\Omega \int_0^t |f_n g_n - fg|\, d\nu_M dP + \int_\Omega \int_0^t |f_n g_n - fg|\, d\nu_N dP\right)$$

and by the choice of the f_n and g_n, these both converge to 0. Similar considerations apply to

$$\left| \int_\Omega \int_0^t f_n g_n d[M-N]\, dP - \int_\Omega \int_0^t fg\, d[M-N]\, dP \right|$$

and show

$$\lim_{n\to\infty} \int_\Omega \int_0^t f_n g_n d[M,N] = \int_\Omega \int_0^t fg\, d[M,N] \ \blacksquare$$

33.1 The Stieltjes Integral

When we do Stieltjes integration, the first thing to consider is continuous functions. I am going to do this here. It seems to me that this is an important case to consider if for no other reason than this is what we do with Stieltjes integration. If M is continuous and f is of bounded variation, and $\left\{t_k^n\right\}_{k=0}^{m_n-1}$ is a partition P_n of $[0,T]$ for which $\|P_n\| \equiv \max\left\{\left|t_{k+1}^n - t_k^n\right|, k=0,1,...,m_n-1\right\}$, then

$$\lim_{n\to\infty} \sum_{k=0}^{m_n-1} f\left(t_k^n\right)\left(M\left(t\wedge t_{k+1}^n\right) - M\left(t\wedge t_k^n\right)\right) = \int_0^t f dM \tag{33.14}$$

exists in U. This is because of the integration by parts theorem for Stieltjes integrals. Indeed, $\int_0^t M df$ exists by standard arguments. This is a little different here because M has values in a Hilbert space, but the proof is essentially the same as for scalar valued functions. See the proof in my single variable advanced calculus book, or [2] or [28]. Thus for a.e. ω, one obtains the limit in 33.14 directly from the theory of Stieltjes integration, this limit taking place in U. Suppose now that

$$M^* \in L^2(\Omega)$$

In the above context, 33.11 is

$$E\left(\left\|\int_0^t f dM\right\|_U^2\right) = \int_\Omega \int_0^t f(s)^2 d[M](s) dP$$

when f is an elementary function. Let $f(t,\omega)$ be bounded and continuous in t with $f(t,\cdot)$ $\mathscr{F}_t$ measurable. Let $P_n = \left\{t_j^n\right\}_{j=0}^{m_n-1}$ and let f_k be the elementary function

$$f_k(t,\omega) \equiv \sum_{j=0}^{m_n-1} f_k\left(t_j^n\right) \mathscr{X}_{(t_j^n, t_{j+1}^n]}(t)$$

Thus these elementary functions converge uniformly to f for $t\in[0,T]$ for fixed ω as $\|P_n\| \to 0$ and they are all bounded uniformly. Therefore, from 33.11 and the maximal estimates of Theorem 30.5.3,

$$\frac{1}{2}E\left(\sup_{t\le T}\left\|\int_0^t f_k dM - \int_0^t f_m dM\right\|_U^2\right) \le$$

$$E\left(\left\|\int_0^T f_k dM - \int_0^T f_m dM\right\|_U^2\right) = \int_\Omega \int_0^T |f_k(s) - f_m(s)|^2 d[M](s) dP \tag{33.15}$$

This is because you can assume, by taking the union of the two partitions involved, that you are dealing with a single partition for both f_k and f_m. Now for a.e. ω,

$$\lim_{k,m\to\infty} \int_0^T |f_k(s) - f_m(s)|^2 d[M](s) = 0.$$

This is because of the Burkholder Davis Gundy inequality which implies $d[M]$ is a finite measure for a.e. ω. Indeed, from this inequality,

$$c\int_\Omega \left(([M](T))^{1/2}\right)^2 dP = c\int_\Omega [M](T) dP$$

$$= c\int_\Omega \int_0^T d[M]\,dP \le \int_\Omega (M^*)^2\,dP < \infty \tag{33.16}$$

Hence for a.e. $\omega, \lim_{k,m\to\infty} \int_0^T |f_k(s) - f_m(s)|^2\, d[M](s) = 0$. Also from the fact these elementary functions are all bounded, that inside integral on the right in 33.15 is no more than $K[M(T)]$ for a constant K which comes from the upper bound of all these elementary functions. This is a function in $L^1(\Omega)$ by 33.16. Now if A is a measurable set,

$$\int_A \int_0^T |f_{n+k}(s) - f_n(s)|^2\, d[M](s)\,dP \le K\int_A [M](T)$$

and $[M](T)$ is a function in L^1 so this collection of functions of ω ,

$$\int_0^T |f_k(s) - f_m(s)|^2\, d[M](s)$$

is uniformly integrable. By the Vitali convergence theorem,

$$\lim_{k,m\to\infty} \int_\Omega \int_0^T |f_k(s) - f_m(s)|^2\, d[M](s)\,dP = 0.$$

By 33.15, $\left\{\int_0^t f_k dM\right\}_{k=1}^\infty$ is a Cauchy sequence in $\mathscr{M}_T^2(U)$ and so there is a unique martingale $I(t) \equiv \int_0^t f dM$ such that $\int_0^t f_k dM \to I(t)$ in $\mathscr{M}_T^2(U)$. Also by Proposition 31.7.2, there is a subsequence which converges uniformly on $[0,T]$ for a.e. ω. In case f is of bounded variation in addition to being continuous because in this case, the Stieltjes sums $\int_0^t f_{n_k} dM$ converge to $\int_0^t f dM$. This proves the following interesting relationship.

Proposition 33.1.1 *In the above context where $M^* \in L^2(\Omega)$, M a martingale with values in U a separable Hilbert space, suppose $t \to f(t,\omega)$ is of bounded variation and is continuous and $\omega \to f(t,\omega)$ is adapted to the filtration $\mathscr{F}_t$. Also suppose $(t,\omega) \to f(t,\omega)$ is bounded. Then the ordinary Stieltjes integral $\int_0^t f(s)\,dM(\omega)$ is a martingale.*

In the above argument, it was not necessary that $t \to f(t,\omega)$ have bounded variation so the $I(t)$ is in a sense more general than the Stieltjes integral but it extends the idea of the Stieltjes integral.

What if M is only a local martingale? Then you could let σ_m be the first hitting time of m by $\|M(t)\|$ and you could repeat everything and get

$$\int_0^t f(s)\,dM^{\sigma_m} = \int_0^{t\wedge\sigma_m} f(s)\,dM$$

is a martingale. The approximating sums in this case would be

$$\sum_{k=0}^{m_n-1} f(t_k^n)\left(M^{\sigma_m}(t\wedge t_{k+1}^n) - M^{\sigma_m}(t\wedge t_k^n)\right) \to \int_0^t f(s)\,dM^{\sigma_m}$$

$$\sum_{k=0}^{m_n-1} f(t_k^n)\left(M^{\sigma_m}(t\wedge t_{k+1}^n) - M^{\sigma_m}(t\wedge t_k^n)\right)$$

$$= \left(\sum_{k=0}^{m_n-1} f(t_k^n)\left(M(t\wedge t_{k+1}^n) - M(t\wedge t_k^n)\right)\right)^{\sigma_m} \to \int_0^{t\wedge\sigma_m} f(s)\,dM$$

as $\|P_n\| \to \infty$. Thus, more generally, this Stieltjes integral $\int_0^t f dM$ is a local martingale. This is all written with scalar valued $f(s,\omega)$ in mind, but if $f(s,\omega)$ were something in $\mathscr{L}(U,U)$ would it be any different? In case $[M]$ depends only on t, all the above considerations become easier. Indeed, if $[M]$ is just an increasing function of t,33.15 would imply that one could get $\int_0^t f_k(s)\,dM$ is a Cauchy sequence in $\mathscr{M}_T^2(U)$ whenever $\lim_{k,m\to\infty} \int_\Omega \int_0^T |f_k(s) - f_m(s)|^2\, dF(s)\, dP = 0$ where $F(s) = [M](s)$ and F an increasing function. In fact, the main interest will be when $[M](t) = t$ as in Example 32.5.7 so $\left\{\int_0^t f_k(s)\,dM\right\}$ being a Cauchy sequence in $\mathscr{M}_T^2(U)$ comes from assuming simply that $\{f_k\}$ is Cauchy in $L^2\left(\Omega; L^2([0,T])\right)$.

33.2 The Stochastic Integral When $f(s) \in \mathscr{L}_2(U,H)$

Let H, U be separable Hilbert spaces and suppose for each s, $f(s) \in \mathscr{L}_2(U,H)$, the space of Hilbert Schmidt operators described in the section on compact operators Section 22.5.2. Recall that $f(s) \in \mathscr{L}_2(U,H)$ implies $f(s)^* f(s)$ is a self adjoint compact operator thanks to Theorem 22.5.18. Thus as pointed out there, we can pick **any** orthonormal basis $\{e_k\}$ for U and

$$\|f^*(s) f(s)\|_{\mathscr{L}_2}^2 = \sum_{k=1}^{\infty} \|f^*(s) f(s) e_k\|_H^2$$

the same value being obtained for any of these orthonormal sets. Now also $f^*(s) f(s)$ is compact and self adjoint so by the Hilbert Schmidt theorem, Theorem 22.5.3, there is a decreasing list of positive numbers $\{\lambda_k\}$ and a corresponding orthonormal set of eigenvectors $\{e_k\}$ such that $f^*(s) f(s) e_k = \lambda_k e_k$. Then from this equation, $\|f(s) e_k\|^2 = \lambda_k$ and so $\sum_k |\lambda_k| < \infty$.

Also, for $L \in \mathscr{L}_2(U,H)$ since L^*L is nonnegative, there is a self adjoint square root $\sqrt{L^*L}$ and $\sqrt{L^*L} e_k = \sqrt{\lambda_k} e_k$. Note also, $\left\|\sqrt{L^*L} e_k\right\|^2 = \|Le_k\|^2$. As pointed out earlier, This implies that for any pair of orthonormal basis $\{e_k\}, \{\hat{e}_k\}$

$$\sum_k \|Le_k\|^2 = \sum_k \left\|\sqrt{L^*L} e_k\right\|^2 = \sum_k \left\|\sqrt{L^*L} \hat{e}_k\right\|^2 = \sum_k \|L\hat{e}_k\|^2$$

so the norm of something in $\mathscr{L}_2(U,H)$ can be defined using any orthonormal basis.

Definition 33.2.1 *An elementary function is one of the form*

$$f(t) \equiv \sum_{r=0}^{m_n - 1} f_r \mathscr{X}_{(t_r, t_{r+1}]}(t)$$

where $f_r \in \mathscr{L}_2(U,H)$ and f_r is $\mathscr{F}_{t_r}$ measurable. In this section, assume also that f_r is bounded. Here $t_0 < t_1 < \cdots < t_{m_n}$ is a partition of $[0,T]$ as above. We can define the stochastic integral of an elementary function with respect to a continuous martingale $M(t)$ as

$$\int_0^t f dM \equiv \sum_{r=0}^{m_n - 1} f_r \left(M(t \wedge t_{r+1}) - M(t \wedge t_r)\right)$$

Note that $\int_0^t f dM \in H$ for a given ω. Also note that, since the $\mathscr{F}_t$ are increasing, given two elementary functions, we can write them both with respect to the same partition and consequently this set of elementary functions is a linear space and the integral on these functions is linear.

Note that for $t_r < s \leq t$,

$$\begin{aligned} E\left(f_r\left(M\left(t\wedge t_{r+1}\right)-M\left(t\wedge t_r\right)\right)|\mathscr{F}_s\right) &= f_r E\left(M\left(t\wedge t_{r+1}\right)-M\left(t\wedge t_r\right)|\mathscr{F}_s\right) \\ &= f_r\left(M\left(s\wedge t_{r+1}\right)-M\left(s\wedge t_r\right)\right) \end{aligned}$$

However, for $t \leq t_r$ the term in the sum equals 0. Thus each term in that sum is a martingale. It follows that $\int_0^t f dM$ is a martingale for f an elementary function. It is also a continuous martingale because each term is continuous. By Theorem 30.5.3, the maximal estimate and the theorem about the quadratic variation,

$$E\left(\sup_{t\in[0,T]}\left\|\int_0^t f dM\right\|^2\right) \leq 2E\left(\left\|\int_0^T f dM\right\|^2\right)$$

Now from the definition of the integral given above, $E\left(\left\|\int_0^T f dM\right\|^2\right) =$

$$E\left(\sum_{r=0}^{m_n-1} f_r\left(M\left(t_{r+1}\right)-M\left(t_r\right)\right), \sum_{r=0}^{m_n-1} f_r\left(M\left(t_{r+1}\right)-M\left(t_r\right)\right)\right)_H \tag{33.17}$$

because $T \geq t_r$ for each t_r in the partition of the interval. Consider a mixed term in the above product in which $j < k$

$$E\left(f_k\left(M\left(t_{k+1}\right)-M\left(t_k\right)\right), f_j\left(M\left(t_j\right)-M\left(t_j\right)\right)\right)$$

$$\begin{aligned} &= E\left(E\left[f_k\left(M\left(t_{k+1}\right)-M\left(t_k\right)\right), f_j\left(M\left(t_j\right)-M\left(t_j\right)\right)|\mathscr{F}_{t_k}\right]\right) \\ &= E\left(E\left[\left(M\left(t_{k+1}\right)-M\left(t_k\right)\right), f_k^* f_j\left(M\left(t_j\right)-M\left(t_j\right)\right)|\mathscr{F}_{t_k}\right]\right) \\ &= E\left(f_k^* f_j\left(M\left(t_j\right)-M\left(t_j\right)\right), E\left[\left(M\left(t_{k+1}\right)-M\left(t_k\right)\right)|\mathscr{F}_{t_k}\right]\right) = 0 \end{aligned}$$

Now from Lemma 22.5.15 on Hilbert Schmidt operators,

$$\|f_r M\|^2 \leq \|f_r\|^2 \|M\|^2 \leq \|f_r\|_{\mathscr{L}_2}^2 \|M\|^2.$$

Therefore, from 33.17,

$$E\left(\sum_{r=0}^{m_n-1} f_r\left(M\left(t_{r+1}\right)-M\left(t_r\right)\right), \sum_{r=0}^{m_n-1} f_r\left(M\left(t_{r+1}\right)-M\left(t_r\right)\right)\right)_H =$$

$$\sum_{r=0}^{m_n-1} E\left(\left\|f_r\left(M\left(t_{r+1}\right)-M\left(t_r\right)\right)\right\|^2\right) \leq \sum_{r=0}^{m_n-1} E\left(\|f_r\|^2\left\|\left(M\left(t_{r+1}\right)-M\left(t_r\right)\right)\right\|^2\right)$$

$$= \sum_{r=0}^{m_n-1} E\left(\|f_r\|_{\mathscr{L}_2}^2 \left\|M\left(t_{r+1}\right)\right\|^2\right)$$

$$+ \sum_{r=0}^{m_n-1} E\left(\|f_r\|_{\mathscr{L}_2}^2 \left\|M\left(t_r\right)\right\|^2\right) - 2\sum_{r=0}^{m_n-1} E\left(\|f_r\|_{\mathscr{L}_2}^2 \left(M\left(t_{r+1}\right), M\left(t_r\right)\right)\right) \tag{33.18}$$

Consider the mixed term on the end.

$$E\left(\|f_r\|_{\mathscr{L}_2}^2 \left(M\left(t_{r+1}\right), M\left(t_r\right)\right)\right) = E\left(E\left(\|f_r\|_{\mathscr{L}_2}^2 \left(M\left(t_{r+1}\right), M\left(t_r\right)\right)|_{\mathscr{F}_{t_r}}\right)\right)$$

$$= E\left(\|f_r\|^2_{\mathscr{L}_2} M(t_r), E\left(M(t_{r+1})\,|_{\mathscr{F}_{t_r}}\right)\right) = E\left(\|f_r\|^2_{\mathscr{L}_2} \|M(t_r)\|^2\right)$$

Thus 33.18 reduces to

$$\begin{aligned} & \sum_{r=0}^{m_n-1} E\left(\|f_r\|^2_{\mathscr{L}_2} \|M(t_{r+1})\|^2\right) - E\left(\|f_r\|^2_{\mathscr{L}_2} \|M(t_r)\|^2\right) \\ = & \int_\Omega \sum_{r=0}^{m_n-1} \|f_r\|^2_{\mathscr{L}_2} \left(\|M(t_{r+1})\|^2 - \|M(t_r)\|^2\right) dP \\ = & \int_\Omega \sum_{r=0}^{m_n-1} \|f_r\|^2_{\mathscr{L}_2} \left([M(t_{r+1})] - [M(t_r)]\right) dP \end{aligned}$$

because the martingales from the quadratic variation have expectation 0. This has proved the following theorem.

Theorem 33.2.2 *Let f be an elementary function corresponding to the partition $P = \{t_k\}_{k=0}^{m_n}$. Then*

$$\begin{aligned} \frac{1}{2} E\left(\sup_{t\in[0,T]} \left\|\int_0^t f dM\right\|^2\right) & \leq \int_\Omega \sum_{r=0}^{m_n-1} \|f_r\|^2_{\mathscr{L}_2} \left([M(t_{r+1})] - [M(t_r)]\right) dP \\ & = \int_\Omega \int_0^T \|f\|^2_{\mathscr{L}_2} d[M](t)\, dP \end{aligned} \tag{33.19}$$

That integral on the right end is just the conventional Stieltjes integral of a step function. Recall that $[M]$ is increasing and continuous.

Theorem 33.2.3 *Let f be uniformly bounded, $\|f(t,\omega)\|_{\mathscr{L}_2(U,H)} \leq K$, continuous in t, and adapted. Also let $M(t)$ be a bounded continuous martingale with values in U and suppose $[M](T) \in L^1(\Omega,P)$. Let f_n be an elementary function approximating f*

$$f_n(t) \equiv \sum_{k=0}^{m_n-1} f(t_k^n)\, \mathscr{X}_{(t_k^n, t_{k+1}^n]}(t)$$

where $P_n = \{t_k^n\}_{k=0}^{m_n}$ is a partition of $[0,T]$. Assume

$$\|P_n\| \equiv \max\left\{\left|t_{k+1}^n - t_k^n\right| : k \leq m_n\right\}.$$

Then there exists a unique continuous bounded martingale denoted as $\int_0^t f dM$ which satisfies

$$\lim_{\|P_n\|\to 0} \int_0^t f_n dM = \int_0^t f dM \text{ in } \mathscr{M}_T^2(H)$$

where this means: For every $\varepsilon > 0$ there is $\delta > 0$ such that if P_n is a partition having $\|P_n\| < \delta$, then

$$\left\|\int_0^{(\cdot)} f_n dM - \int_0^{(\cdot)} f dM\right\|_{\mathscr{M}_T^2(H)} < \varepsilon.$$

For any such sequence of partitions and approximating elementary functions, there is a set of measure zero N such that if $\omega \notin N$, then $\int_0^t f_n dM(\omega) \to \int_0^t f dM(\omega)$ uniformly in $t \in [0,T]$. Also, for such f just described,

$$\frac{1}{2} E\left(\sup_{t\in[0,T]} \left\|\int_0^t f dM\right\|^2\right) \leq \int_\Omega \int_0^T \|f\|^2_{\mathscr{L}_2} d[M](t)\, dP \tag{33.20}$$

Proof: Letting f_n be an approximating elementary function for f corresponding to P_n a partition with $\|P_n\| \to 0$, it follows that for each $\omega, f_n(t) \to f(t)$ uniformly in t. From 33.19,

$$\frac{1}{2}\left\|\int_0^t f_n dM(\omega) - \int_0^t f_m dM(\omega)\right\|^2_{\mathscr{M}_T^2(H)} \leq \int_\Omega \int_0^T \|f_n - f_m\|^2_{\mathscr{L}_2} d[M](t) dP \tag{33.21}$$

By the uniform convergence to f, the inside Stieltjes integral converges to 0 as $n,m \to \infty$. The integrand is bounded by $4K^2[M](T) \in L^1(\Omega)$ and so one can apply the dominated convergence theorem to conclude that $\left\{\int_0^t f_n dM(\omega)\right\}$ is a Cauchy sequence in $\mathscr{M}_T^2(H)$. Therefore, there is a unique $\int_0^t f dM \in \mathscr{M}_T^2(H)$ to which these $\int_0^t f_n dM$ converge. By Proposition 31.7.2 there exists a subsequence converging uniformly in t for all ω off some set of measure zero. The above inequality in 33.21 also implies that $\int_0^{(\cdot)} f dM$ is independent of approximating sequence of elementary functions.

Now $\frac{1}{2}E\left(\sup_{t\in[0,T]}\left\|\int_0^t f_n dM\right\|^2\right) \leq \int_\Omega \int_0^T \|f_n\|^2_{\mathscr{L}_2} d[M](t) dP$ and so, passing to a limit using the dominated convergence theorem yields 33.20. ■

Definition 33.2.4 *Let f be continuous in t, adapted, and uniformly bounded, and let M be a continuous bounded martingale. Then one can define for ω off a set of measure 0, $\int_0^t f dM$ where*

$$E\left(\sup_{t\in[0,T]}\left\|\int_0^t f dM - \int_0^t f_n dM\right\|^2\right) = 0 \tag{33.22}$$

where $\{f_n\}$ is any sequence of elementary functions converging to f.

If f is of bounded variation in addition to being continuous, then the convergence for each ω follows directly from considerations involving Stieltjes integrals without any probabilistic considerations. However, here f is only required to be continuous. Later, this will be relaxed further.

Note that there was no need for f to have values in $\mathscr{L}_2(U,H)$. It would suffice to have $f \in \mathscr{L}(U,H)$ and all of the above would work the same way. You would simply replace $\|f\|_{\mathscr{L}_2(U,H)}$ with $\|f\|_{\mathscr{L}(U,H)}$. The reason for the specialization involves some technical considerations relative to the Wiener process and the need for compactness. Therefore, I have chosen to present it from the beginning with this more specialized case.

Now here is a fundamental lemma about integrals of these elementary functions having to do with stopping times.

Lemma 33.2.5 *Suppose M is a continuous bounded martingale having values in a separable Hilbert space U and let f be an elementary function having values in $\mathscr{L}_2(U,H)$ for H a separable Hilbert space. Then the above inequality 33.20 is valid. Also, if σ is a stopping time, then*

$$\int_0^t f dM^\sigma = \int_0^{t\wedge\sigma} f dM \tag{33.23}$$

Also,

$$\frac{1}{2}E\left(\sup_{t\in[0,T]}\left\|\int_0^{t\wedge\sigma} f dM\right\|^2\right) \leq \int_\Omega \int_0^T \left\|f\mathscr{X}_{[0,\sigma]}\right\|^2_{\mathscr{L}_2} d[M] dP \tag{33.24}$$

Proof: It only remains to verify 33.23. The first equal sign is obvious from the definition of $\int_0^t f dM$. Both sides equal

$$\sum_{r=0}^{m_n-1} f_r \left(M\left(t \wedge \sigma \wedge t_{r+1}\right) - M\left(t \wedge \sigma \wedge t_r\right)\right)$$

Now consider 33.24. From 33.23, the left side is

$$\begin{aligned} \frac{1}{2} E\left(\sup_{t\in[0,T]} \left\| \int_0^t f dM^\sigma \right\|^2 \right) &\leq \int_\Omega \int_0^T \|f\|^2_{\mathscr{L}_2} d\left[M^\sigma\right] dP \\ &= \int_\Omega \int_0^T \|f\|^2_{\mathscr{L}_2} d\left[M\right]^\sigma dP \end{aligned}$$

$$= \int_\Omega \int_0^T \mathscr{X}_{[0,\sigma]} \|f\|^2_{\mathscr{L}_2} d\left[M\right] dP = \int_\Omega \int_0^T \left\| \mathscr{X}_{[0,\sigma]} f \right\|^2_{\mathscr{L}_2} d\left[M\right] dP$$

This is because when $t > \sigma$, $[M]^\sigma(t) = [M](\sigma)$, a constant. Thus the contribution to the conventional integral is 0 from then on. On the other hand, $[M]^\sigma(t) = [M](t)$ for $t \leq \sigma$ and so the last equation follows. ■

Now here is a nice proposition which is a summary of what has just been discussed along with some other observations.

Proposition 33.2.6 *For $\|f(t,\omega)\|$ bounded by K and continuous in t for each ω having values in $\mathscr{L}_2(U,H)$, and for M a continuous bounded martingale with values in U, $t \to \int_0^t f dM$ is a continuous martingale with values in H. Assume that $[M](T) \in L^1(\Omega)$. Also the fundamental inequality*

$$\frac{1}{2} E\left(\sup_{t\in[0,T]} \left\| \int_0^t f dM \right\|^2 \right) \leq \int_\Omega \int_0^T \|f\|^2_{\mathscr{L}_2} d\left[M\right] dP \tag{33.25}$$

is valid for f. In addition, $f \to \int_0^t f dM$ is linear on the linear space of bounded continuous in t adapted functions. If σ is a stopping time,

$$\int_0^{t\wedge\sigma} f dM = \int_0^t f dM^\sigma \tag{33.26}$$

Proof: Since the integral is a limit of integrals of elementary functions for ω off a set of measure zero and since this integral is linear on these functions, the integral is linear. Finally, consider the claim 33.26. From Proposition 31.7.2, and letting $\{f_n\}$ be elementary functions approximating f as above, then by that proposition again, there is a further subsequence still denoted with n such that for a.e. ω,

$$\int_0^t f dM^\sigma = \lim_{n\to\infty} \int_0^t f_n dM^\sigma = \lim_{n\to\infty} \int_0^{t\wedge\sigma} f_n dM = \int_0^{t\wedge\sigma} f dM \blacksquare$$

Note how this does not require f to be of bounded variation. If f were of bounded variation, you would get pointwise convergence of the Stieltjes sums for $\int_0^t f_n dM$ to $\int_0^t f dM$ as a consequence of simple considerations involving Stieltjes integrals. Then the new information is that the Stieltjes integral $\int_0^t f dM$ for f of bounded variation is a continuous martingale.

33.3 More on Stopping Times

Next I want to consider $\int_0^t f\mathscr{X}_{[0,\sigma]}dM$ in the case of elementary functions for f a continuous in t, adapted, and bounded. This involves the same process as earlier and it is analogous to the traditional definition of the Stieltjes integral. First we approximate with an elementary function and then pass to a limit.

Proposition 33.3.1 *Let σ be a stopping time and let $\{t_k\}_{k=0}^{m_n}$ be partition points of $[0,T]$. Also define the discrete approximation of σ*

$$\sigma_n(\omega) \equiv \sum_{k=0}^{m_n-1} t_k \mathscr{X}_{\sigma^{-1}((t_k,t_{k+1}])}(\omega)$$

Then for f an elementary function with respect to $\{t_k\}_{k=0}^{n}$ points as described above, it follows that $f\mathscr{X}_{[0,\sigma_n]}$ is an elementary function and σ_n is a stopping time. Also

$$\int_0^t f\mathscr{X}_{[0,\sigma_n]}dM = \int_0^{t\wedge\sigma_n} fdM$$

Proof: First, why is σ_n a stopping time? Consider $[\sigma_n \le t]$. Say $t \in (t_k,t_{k+1}]$. In case, $t = t_{k+1}$, $[\sigma_n \le t] = [\sigma \le t] \in \mathscr{F}_t$. Otherwise, $[\sigma_n \le t] = [\sigma \le t_k] \in \mathscr{F}_{t_k} \subseteq \mathscr{F}_t$. Thus σ_n is indeed a stopping time. Now

$$f\mathscr{X}_{[0,\sigma_n]}(t) = \sum_{k=0}^{m_n-1} f_k \mathscr{X}_{[0,\sigma_n]}(t)\,\mathscr{X}_{(t_k,t_{k+1}]}(t).$$

t is somewhere. Say $t \in (t_k,t_{k+1}]$. Then consider $f_k\mathscr{X}_{[0,\sigma_n]}(t)$. For this t, this term is nonzero if and only if $\omega \in [t \le \sigma_n]$ if and only if $\omega \in [t_k < \sigma_n] \in \mathscr{F}_{t_k}$. Thus this term is the indicator function of a set in $\mathscr{F}_{t_k}$ for all $t \in (t_k,t_{k+1}]$. It follows $f\mathscr{X}_{[0,\sigma_n]}$ is an elementary function and can be written as $\sum_{k=0}^{m_n-1} f_k\mathscr{X}_{[t_k<\sigma_n]}\mathscr{X}_{(t_k,t_{k+1}]}(t)$, so its integral is

$$\begin{aligned} &\sum_{k=0}^{m_n-1} f_k\mathscr{X}_{[t_k<\sigma_n]}(M(t\wedge t_{k+1}) - M(t\wedge t_k)) \\ = &\sum_{k=0}^{m_n-1} f_k(M(t\wedge t_{k+1}\wedge\sigma_n) - M(t\wedge t_k\wedge\sigma_n)) \end{aligned}$$

because if $\sigma_n > t_k$ in the k^{th} term and the term is nonzero, then $\sigma_n = t_{k+1}$. If $\sigma_n \le t_k$, the k^{th} term on the left is 0 and on the right that term is

$$f_k(M(t\wedge\sigma_n) - M(t\wedge\sigma_n)),$$

also zero. Now the right side in the above is just $\int_0^{t\wedge\sigma_n} fdM$ and the left side is defined earlier as $\int_0^t f\mathscr{X}_{[0,\sigma_n]}dM$. ■

Lemma 33.3.2 *Let f be bounded, adapted, and continuous in t and let σ be a stopping time with finite values in $[0,T]$. Also assume $[M](T) \in L^1(\Omega,P)$. Then letting σ_n be as above,*

$$\sigma_n(\omega) \equiv \sum_{k=0}^{m_n-1} t_k \mathscr{X}_{\sigma^{-1}((t_k,t_{k+1}])}(\omega),$$

$f_n\mathscr{X}_{[0,\sigma_n]} \to f\mathscr{X}_{[0,\sigma]}$ *in* $L^2(\Omega\times[0,T];\mathscr{L}_2(U,H))$. *Here* $f_n(t)$ *is the elementary function* $\sum_{k=0}^{m_n-1} f(t_k^n)\,\mathscr{X}_{(t_k^n,t_{k+1}^n]}(t)$ *where* $\{t_k^{m_n}\}$ *is a partition* P_n *with* $\|P_n\|\to 0$ *so that* f_n *converges uniformly to* f *for each* ω.

Proof: Note that $|\sigma_n(\omega)-\sigma(\omega)|\le\|P_n\|$. Then

$$\begin{aligned}\left\|f_n\mathscr{X}_{[0,\sigma_n]}-f\mathscr{X}_{[0,\sigma]}\right\| &\le \left\|(f_n-f)\,\mathscr{X}_{[0,\sigma_n]}\right\|+\left\|f\mathscr{X}_{[\sigma_n,\sigma]}\right\|\\ \left\|f_n\mathscr{X}_{[0,\sigma_n]}-f\mathscr{X}_{[0,\sigma]}\right\|^2 &\le 2\left(\left\|(f_n-f)\,\mathscr{X}_{[0,\sigma_n]}\right\|^2+\left\|f\mathscr{X}_{[\sigma_n,\sigma]}\right\|^2\right)\end{aligned}$$

Thus

$$\begin{aligned}&\int_\Omega\int_0^T\left\|f_n\mathscr{X}_{[0,\sigma_n]}-f\mathscr{X}_{[0,\sigma]}\right\|^2 d[M]\,dP\\ \le\ & 2\int_\Omega\int_0^T\|f_n-f\|^2 d[M]\,dP+2K\int_\Omega\int_0^T\mathscr{X}_{[\sigma_n,\sigma]}d[M]\,dP\\ \le\ & 2\int_\Omega\int_0^T\|f_n-f\|^2 d[M]\,dP+2K\int_\Omega[M](\sigma)-[M](\sigma_n)\,dP \qquad (33.27)\end{aligned}$$

the integrands $\int_0^T\|f_n-f\|^2 d[M]$ and $[M](\sigma)-[M](\sigma_n)$ both converge to 0 a.e. ω as $n\to\infty$ thanks to continuity of $[M]$. Also the assumption that $[M](T)$ is in L^1 along with the boundedness of f imply these integrands are uniformly integrable. Hence we can use the Vitali convergence theorem and conclude that the limit of 33.27 is 0. ■

Now note that if you fix $\omega, f_n\mathscr{X}_{[0,\sigma_n]}(t)\to f\mathscr{X}_{[0,\sigma]}(t)$ for each $t<\sigma(\omega)$. Thus, if f is of bounded variation as well as being continuous in each t, standard Stieltjes integral considerations involving continuity of f and M show that $\int_0^t f_n\mathscr{X}_{[0,\sigma_n]}(t)\,dM\to\int_0^t f\mathscr{X}_{[0,\sigma]}(t)\,dM$ with no probabilistic complications at all.

Definition 33.3.3 *Let* f *be adapted, continuous in* t, *and bounded. Let* M *be a continuous bounded martingale. Also let* σ *be a stopping time and let* σ_n *be the discreet approximation above relative to partitions* $P_n=\{t_k^n\}_{k=0}^{m_n}$ *where* $\|P_n\|\to 0$ *and* $\{f_n\}$ *the sequence of elementary functions approximating* f,

$$f_n(t)\equiv\sum_{k=0}^{m_n-1} f(t_k^n)\,\mathscr{X}_{(t_k^n,t_{k+1}^n]}(t),$$

Then there exists a martingale denoted as $\int_0^t f\mathscr{X}_{[0,\sigma]}dM$ *such that*

$$\int_0^t f\mathscr{X}_{[0,\sigma]}dM=\lim_{n\to\infty}\int_0^t f_n\mathscr{X}_{[0,\sigma_n]}dM \text{ in } M_T^2(H)$$

This is something new. Earlier we had $\int_0^t f dM$ defined where f is continuous on $[0,T]$. We also have $\int_0^{t\wedge\sigma} f dM$ defined where σ is a stopping time and f is continuous on $[0,T]$. However, $f\mathscr{X}_{[0,\sigma]}$ is not necessarily continuous on $[0,T]$. It is continuous on $[0,T\wedge\sigma]$ so the time intervals are changing as a function of ω.

To begin with, we can stop the martingale $\int_0^t f_n dM$ with the stopping time σ, this denoted as $\int_0^{t\wedge\sigma} f_n dM$. Then

$$\begin{aligned}&\int_0^{t\wedge\sigma} f_n dM-\int_0^{t\wedge\sigma_n} f_n dM\\ =\ &\sum_{k=0}^{m_n-1} f(t_k^n)\begin{pmatrix}(M(t\wedge\sigma\wedge t_{k+1})-M(t\wedge\sigma\wedge t_k))\\ -(M(t\wedge\sigma_n\wedge t_{k+1})-M(t\wedge\sigma_n\wedge t_k))\end{pmatrix}\end{aligned}$$

$$= \sum_{k=0}^{m_n-1} f(t_k^n) \begin{pmatrix} (M(t\wedge\sigma\wedge t_{k+1}) - M(t\wedge\sigma_n\wedge t_{k+1})) \\ -(M(t\wedge\sigma\wedge t_k) - (M(t\wedge\sigma_n\wedge t_k))) \end{pmatrix}$$

Now from maximal theorems,

$$\begin{aligned} & E\left(\sup_{t\in[0,T]}\left\|\int_0^{t\wedge\sigma} f_n dM - \int_0^{t\wedge\sigma_n} f_n dM\right\|^2\right) \\ \leq\ & E\left(\left\|\int_0^{T\wedge\sigma} f_n dM - \int_0^{T\wedge\sigma_n} f_n dM\right\|^2\right) \end{aligned}$$

$$= E\begin{pmatrix} \sum_{k=0}^{m_n-1} \left\|f(t_k^n)(M(\sigma\wedge t_{k+1}) - M(\sigma_n\wedge t_{k+1}))\right\|^2 \\ +\left\|f(t_k^n)(M(\sigma\wedge t_k) - (M(\sigma_n\wedge t_k)))\right\|^2 \end{pmatrix} + ?? \tag{33.28}$$

where ?? is -2 times the expectation of a sum of mixed terms which can be written in the following form after noticing that $M(\sigma\wedge t_k) - (M(\sigma_n\wedge t_k))$ is $\mathscr{F}_{t_k}$ measurable.

$$E\begin{pmatrix} f(t_k^n)^* f(t_k^n) M(\sigma\wedge t_k) \\ -(M(\sigma_n\wedge t_k)), E\left(M(\sigma\wedge t_{k+1}) - M(\sigma_n\wedge t_{k+1})\,|\mathscr{F}_{t_k}\right) \end{pmatrix}$$

$$\begin{aligned} &= E\left(f(t_k^n)^* f(t_k^n) M(\sigma\wedge t_k) - (M(\sigma_n\wedge t_k)), M(\sigma\wedge t_k) - M(\sigma_n\wedge t_k)\right) \\ &= E\left(\left\|f(t_k^n)(M(\sigma\wedge t_k) - M(\sigma_n\wedge t_k))\right\|^2\right) \end{aligned}$$

Therefore 33.28 reduces to

$$\begin{aligned} & E\begin{pmatrix} \sum_{k=0}^{m_n-1} \left\|f(t_k^n)(M(\sigma\wedge t_{k+1}) - M(\sigma_n\wedge t_{k+1}))\right\|^2 \\ -\left\|f(t_k^n)(M(\sigma\wedge t_k) - (M(\sigma_n\wedge t_k)))\right\|^2 \end{pmatrix} \\ =\ & E\left(\left\|f(T)(M(T\wedge\sigma) - M(T\wedge\sigma_n))\right\|^2\right) \end{aligned}$$

This converges to 0 because the integrand converges to 0 since $\sigma_n(\omega)\to\sigma(\omega)$ and the integrand is uniformly bounded by assumption. Thus this has shown that

$$E\left(\sup_{t\in[0,T]}\left\|\int_0^{t\wedge\sigma} f_n dM - \int_0^{t\wedge\sigma_n} f_n dM\right\|^2\right)\to 0.$$

This has shown the following technical lemma.

Lemma 33.3.4 *Letting f be bounded and continuous in t and adapted and letting M be a bounded continuous martingale, and σ_n the discrete approximation of a stopping time σ and f_n the elementary function approximating f as described above, then it follows that*

$$E\left(\sup_{t\in[0,T]}\left\|\int_0^{t\wedge\sigma} f_n dM - \int_0^{t\wedge\sigma_n} f_n dM\right\|^2\right)\to 0.$$

Theorem 33.3.5 *Definition 33.3.3, is well defined and $\int_0^t f\mathscr{X}_{[0,\sigma]}dM$ is a martingale equal to $\int_0^{t\wedge\sigma} f dM$. Also, there is a set of measure zero N and a subsequence, still denoted as n such that for $\omega\notin N$,*

$$\int_0^t f_n\mathscr{X}_{[0,\sigma_n]}dM(\omega)\to\int_0^t f\mathscr{X}_{[0,\sigma]}dM(\omega) \tag{33.29}$$

uniformly in $t\in[0,T]$.

Proof: From Proposition 31.7.2, $\int_0^{t\wedge\sigma} f_n dM \to \int_0^{t\wedge\sigma} f dM$ in $\mathscr{M}_T^2(H)$. From Lemma 33.3.2 it follows that also

$$\int_0^{t\wedge\sigma_n} f_n dM = \int_0^t f_n \mathscr{X}_{[0,\sigma_n]} dM \to \int_0^{t\wedge\sigma} f dM$$

in $\mathscr{M}_T^2(H)$. Therefore, $\left\{\int_0^t f_n \mathscr{X}_{[0,\sigma_n]} dM\right\}$ is a Cauchy sequence in $\mathscr{M}_T^2(H)$ and so converges to a continuous martingale denoted as $\int_0^t f\mathscr{X}_{[0,\sigma]} dM$. Any two sequences of approximating elementary functions lead to the same $\int_0^t f\mathscr{X}_{[0,\sigma]} dM$ Thanks to Lemma 33.3.2 and the inequality

$$\left\| \int_0^{t\wedge\sigma} f_n dM - \int_0^{t\wedge\sigma} \hat{f}_n dM \right\|^2_{\mathscr{M}_T^2(H)} \leq 2\int_\Omega \int_0^T \left\| f_n - \hat{f}_n \right\|^2_{\mathscr{L}_2} d[M]\, dP$$

in which the right side converges to 0. By Proposition 31.7.2 again, there is a set of measure zero N such that if $\omega \notin N$, then 33.29 holds. ■

Again, if f is not just continuous and adapted but is also of bounded variation, the convergence for each ω follows from Stieltjes integration theory. Essentially, what the above shows is that even in this case, the integral is a continuous martingale. Also, the above gives meaning to the expression $\int_0^t f\mathscr{X}_{[0,\sigma]} dM(\omega)$ as a limit in $\mathscr{M}_T^2(H)$ of integrals of appropriate elementary functions just as $\int_0^t f dM(\omega)$ was a limit of integrals of appropriate elementary functions. It is the same thing you see with Stieltjes integration but here it is much more elaborate. In place of bounded variation you have adapted and the integrator function is now a martingale.

33.4 Local Martingales as Integrators

Now suppose M is a local martingale. This means that there is a localizing sequence of stopping times $\{\tau_n\}, \tau_n \to \infty$, such that M^{τ_n} is a martingale. Recall Proposition 32.3.3 which says that we can always assume the localizing sequence τ_n makes M^{τ_n} uniformly bounded as well as a martingale.

I will use this fact whenever convenient from now on. We know that $\int_0^{t\wedge\tau} f dM \equiv \int_0^t f dM^\tau$ if M is a bounded martingale so the followng definition will simply extend this idea to give a definition for the stochastic integral in the case where M is continuous but maybe not bounded.

Definition 33.4.1 *Let M be a continuous local martingale with a localizing sequence τ_n for which M^{τ_n} is a bounded martingale. Let f be bounded and continuous in t and adapted. Then $\int_0^t f dM$ is defined as follows.*

$$\int_0^{t\wedge\tau_n} f dM \equiv \int_0^t f dM^{\tau_n}$$

thus for a given ω, $\int_0^t f dM = \lim_{n\to\infty} \int_0^t f dM^{\tau_n}$.

Lemma 33.4.2 *The above definition does define an adapted process $\int_0^t f dM$ which is a continuous local martingale.*

Proof: First note that if σ_n is another localizing sequence, $\int_0^t f dM(\omega)$ is the same when defined from either localizing sequence. The reason is that for either sequence, there is n

such that if $m \geq n$, then for a given ω, $\tau_m(\omega), \sigma_m(\omega)$ are both ∞. It follows upon using approximation with elementary functions and passing to a limit using a suitable subsequence of elementary functions that for such τ_m, σ_m,

$$\int_0^{t\wedge\tau_m} f dM(\omega), \int_0^{t\wedge\sigma_m} f dM(\omega) = \lim_{p\to\infty} \sum_{k=0}^{p-1} f\left(t_k^p\right)\left(M\left(t\wedge t_{k+1}^p\right) - M\left(t\wedge t_k^p\right)\right)(\omega)$$

this limit being independent of the localizing sequence used. The reason it is a local continuous martingale is that M^{τ_n} is a bounded continuous martingale and so

$$\left(\int_0^t f dM\right)^{\tau_n} \equiv \int_0^{t\wedge\tau_n} f dM \equiv \int_0^t f dM^{\tau_n}$$

is a martingale. ■

The idea is that if you know it at $t\wedge\tau_n$ for all τ_n where $\tau_n \to \infty$, then you know it at t because you can simply pick τ_n larger than t and from the above, it doesn't matter which localizing sequence you use.

Now suppose M is just a local martingale so there is a localizing sequence of stopping times $\{\tau_n\}$ such that M^{τ_n} is a bounded martingale and suppose $[M](T) \in L^1(\Omega)$. Let f_n be elementary functions converging to f a bounded, adapted, continuous in t function, convergence uniform in t for each ω. Then from the fundamental inequality above and using the formulas for stopping times,

$$\frac{1}{2}E\left(\sup_{t\in[0,T]}\left\|\int_0^{t\wedge\tau_p} f_n dM\right\|^2\right) = \frac{1}{2}E\left(\sup_{t\in[0,T]}\left\|\int_0^t f_n dM^{\tau_p}\right\|^2\right)$$

$$\leq \int_\Omega\int_0^T \|f_n\|^2 d[M]^{\tau_p}(t)\,dP \leq \int_\Omega\int_0^T \|f_n\|^2 d[M](t)\,dP$$

Since $\|f\|$ is assumed bounded and $[M](T) \in L^1$, it follows that there is a dominating function on the right, namely $K^2\,[M](T)$where $K \geq \|f(t,\omega)\|$ for all (t,ω). Let $n\to\infty$ and use the dominated convergence on the right and either Fatou's lemma or monotone convergence theorem on the left to obtain

$$\frac{1}{2}E\left(\sup_{t\in[0,T]}\left\|\int_0^t f dM\right\|^2\right) \leq \int_\Omega\int_0^T \|f\|^2 d[M](t)\,dP$$

The reason the monotone convergence theorem applies is that on the left, the stopping time effectively restricts the values of t over which the sup is taken until $\tau_p \geq t$. You could also apply Fatou's lemma.

This has shown the following proposition.

Proposition 33.4.3 *Let f be adapted, continuous in t and bounded. Also let M be a local martingale with $[M](T) \in L^1(\Omega)$. Then*

$$\frac{1}{2}E\left(\sup_{t\in[0,T]}\left\|\int_0^t f dM\right\|^2\right) \leq \int_\Omega\int_0^T \|f\|^2 d[M](t)\,dP$$

If f is bounded, continuous, and of bounded variation, and adapted and M is a local martingale, this shows that the Stieltjes integral $\int_0^t f dM$ is a local martingale.

It is clear that if M is a local martingale with localizing sequence τ_p and if σ is a stopping time, then M^σ is also a local martingale with localizing sequence τ_p because $(M^\sigma)^{\tau_p} = (M^{\tau_p})^\sigma$, the latter being a bounded martingale. Now for f continuous in t and bounded and adapted,

$$\int_0^{t\wedge\sigma\wedge\tau_p} f dM \equiv \int_0^t f dM^{\sigma\wedge\tau_p} \equiv \int_0^{t\wedge\tau_p} f dM^\sigma$$

Therefore, from the definition, whenever M is a local martingale and f bounded and continuous in t and adapted,

$$\int_0^{t\wedge\sigma} f dM = \int_0^t f dM^\sigma \tag{33.30}$$

33.5 The Stochastic Integral and the Quadratic Variation

In this simple case of the above, you have a bounded martingale M with values in U and you have $f \in U'$. Thus $f \in \mathscr{L}(U,\mathbb{R})$. Is f actually in $\mathscr{L}_2(U,\mathbb{R})$? By Riesz representation theorem, there is $x \in U$ such that $Rx = f$. Then if $\{g_k\}$ is an orthonormal basis for $U, \sum_k |f(g_k)|^2 = \sum_k |(g_k, Rx)|^2 = \sum_k |(g_k, x)|^2 = \|x\|^2$ because it is just the sum of the squares of the Fourier coefficients of x. Thus $f \in \mathscr{L}_2(U,\mathbb{R})$.

Now an example of a continuous in t, adapted, bounded function in $\mathscr{L}_2(U,\mathbb{R})$ is just $RM(t) \equiv f(t)$. Therefore, it makes perfect sense to consider $\int_0^t (RM)\, dM$. Let $\|P_n\| \to 0$ and let the stochastic integral of an elementary function $f_n(t) = \sum_{k=0}^{m_n-1} RM\left(t_k^n\right) \mathscr{X}_{(t_k^n, t_{k+1}^n]}(t)$ be of the form

$$\sum_{k=0}^{m_n-1} RM(t_k^n)\left(M\left(t\wedge t_{k+1}^n\right) - M\left(t\wedge t_k^n\right)\right)$$

where $\left\{t_k^n\right\}_{k=0}^{m_n}$ is this partition P_n.

ALWAYS assume in this that the partitions are nested, $P_n \subseteq P_{n+1}$.

$$\begin{aligned} Q_n(t) &\equiv \sum_{k=0}^{m_n-1} \left\|M\left(t\wedge t_{k+1}^n\right) - M\left(t\wedge t_k^n\right)\right\|_U^2 \\ &= \sum_{k=0}^{m_n-1} \left\|M\left(t\wedge t_{k+1}^n\right)\right\|^2 + \left\|M\left(t\wedge t_k^n\right)\right\|^2 - 2\left(M\left(t\wedge t_{k+1}^n\right), M\left(t\wedge t_k^n\right)\right) \end{aligned}$$

$$= \sum_{k=0}^{m_n-1} \left\|M\left(t\wedge t_{k+1}^n\right)\right\|^2 - \left\|M\left(t\wedge t_k^n\right)\right\|^2 - 2\left(M\left(t\wedge t_k^n\right), M\left(t\wedge t_{k+1}^n\right) - M\left(t\wedge t_k^n\right)\right)$$

$$= \|M(t)\|^2 - 2\int_0^t (RM_n)\, dM$$

Then passing to a limit, then $Q_n(t) \to Q(t)$ in $L^2(\Omega)$ because $2\int_0^t (RM_n)\, dM$ converges in $\mathscr{M}_T^2(\mathbb{R})$. Using a subsequence, we can also get uniform convergence in t for all ω off a set of measure zero. Thus Q is increasing. It follows

$$Q(t) = \|M(t)\|^2 - 2\int_0^t (RM)\, dM = [M](t) + N(t) - 2\int_0^t (RM)\, dM$$

and so $Q(t)-[M](t)$ equals a martingale. Thus from Lemma 32.2.1, $Q(t)-[M](t)=0$. This proves the first part of the following important result.

Theorem 33.5.1 *Let H be a Hilbert space and suppose $(M,\mathcal{F}_t)$, $t\in[0,T]$ is a uniformly bounded continuous martingale with values in H. Also let $\{t_k^n\}_{k=1}^{m_n}$ be a sequence of partitions satisfying*

$$\lim_{n\to\infty}\max\left\{\left|t_i^n-t_{i+1}^n\right|,i=0,\cdots,m_n\right\}=0,\ \{t_k^n\}_{k=1}^{m_n}\subseteq\left\{t_k^{n+1}\right\}_{k=1}^{m_{n+1}}.$$

Then

$$[M](t)=\lim_{n\to\infty}\sum_{k=0}^{m_n-1}\left\|M\left(t\wedge t_{k+1}^n\right)-M\left(t\wedge t_k^n\right)\right\|_H^2$$

the limit taking place in $L^2(\Omega)$. In case M is just a continuous local martingale, the above limit happens in probability.

Proof: It only remains to show the claim about the case where M is a local martingale. Suppose M is only a continuous local martingale. By Proposition 32.3.3 there exists an increasing localizing sequence $\{\tau_k\}$ such that M^{τ_k} is a uniformly bounded martingale. Then

$$P\left(\cup_{k=1}^{\infty}[\tau_k=\infty]\right)=1$$

As above, let

$$Q_n(t)\equiv\sum_{k=0}^{m_n-1}\left\|M\left(t\wedge t_{k+1}^n\right)-M\left(t\wedge t_k^n\right)\right\|_H^2$$

where there are m_n points in P_n where as before, $P_n\subseteq P_{n+1}$ for all n.

Let $\eta,\varepsilon>0$ be given. Then there exists k large enough that $P([\tau_k=\infty])>1-\eta/2$. This is because the sets $[\tau_k=\infty]$ increase to Ω other than a set of measure zero. Then for this k,

$$\left[\left|Q_n^{\tau_k}-[M]^{\tau_k}(t)\right|>\varepsilon\right]\cap[\tau_k=\infty]=\left[|Q_n-[M](t)|>\varepsilon\right]\cap[\tau_k=\infty]$$

Thus

$$\begin{aligned}P([|Q_n-[M](t)|>\varepsilon]) &\leq P([|Q_n-[M](t)|>\varepsilon]\cap[\tau_k=\infty])\\ &\quad +P([\tau_k<\infty])\\ &\leq P\left(\left[\left|Q_n^{\tau_k}-[M]^{\tau_k}(t)\right|>\varepsilon\right]\right)+\eta/2\end{aligned}$$

The convergence in probability of $Q_n^{\tau_k}(t)$ to $[M]^{\tau_k}(t)$ follows from the convergence in $L^2(\Omega)$ shown earlier for bounded martingales, and so if n is large enough, the right side of the above inequality is less than $\eta/2+\eta/2=\eta$. Since η was arbitrary, this proves convergence in probability. ■

33.6 The Case of $f\in L^2(\Omega\times[0,T];\mathcal{L}_2(U,H))$

At this point I will discontinue the general treatment in terms of arbitrary martingales and suppose that $[M](t)=F(t)$ a continuous increasing function which depends only on t and not on ω. In fact, the most interest is centered on the Wiener process in which $[M](t)=at$ for $a>0$ and this is the case considered here. Of course a does not matter so we will simply assume $[M](t)=t$.

It is not necessary to have f be uniformly bounded. Instead, one can consider $f \in L^2(\Omega \times [0,T]; \mathscr{L}_2(U,H))$ where f is progressively measurable. I considered the case where f is continuous in t above because it is a convenient way to tie this in to the ordinary theory of Stieltjes integrals and to point out that these standard objects do deliver martingales in some reasonable cases.

As before, I will first consider the case where M is a bounded martingale and then extend to the case where M is a local martingale. As before, it is all based on the fundamental inequality

$$\frac{1}{2}E\left(\sup_{t\in[0,T]}\left\|\int_0^t f dM\right\|^2\right) \leq E\left(\left\|\int_0^T f dM\right\|^2\right) \leq \int_\Omega \int_0^T \|f\|^2_{\mathscr{L}_2} dt dP \tag{33.31}$$

which was shown to hold for all adapted f continuous in t and also bounded, which implies the integral on the right is finite.

Let $f \in L^2(\Omega \times [0,T]; \mathscr{L}_2(U,H))$ be adapted and let $f(t,\omega)$ be extended as 0 for t off $[0,T]$.

Definition 33.6.1 *Let* $f_n(t) \equiv n\int_{t-1/n}^t f(s)\,ds$.

Lemma 33.6.2 *f_n is adapted and $t \to f_n(t)$ is continuous for a.e. ω.*

Proof: First consider the claim about continuity. For each ω off a set of measure zero, $f \in L^2([0,T]; \mathscr{L}_2)$ and so, for such ω

$$\begin{aligned}\|f_n(t) - f_n(\hat{t})\|^2_{\mathscr{L}_2} &= \left\|n\int_{t-1/n}^t f(s)\,ds - n\int_{\hat{t}-1/n}^{\hat{t}} f(s)\,ds\right\|^2 dP \\ &= n\left\|\int_{t-1/n}^{\hat{t}-1/n} f(s)\,ds + \int_t^{\hat{t}} f(s)\,ds\right\|^2 dP \\ \leq 2n\left(\left\|\int_{t-1/n}^{\hat{t}-1/n} f(s)\,ds\right\|^2 + \left\|\int_t^{\hat{t}} f(s)\,ds\right\|^2\right) &\leq 8n(t-\hat{t})\|f\|^2_{L^2([0,T];\mathscr{L}_2)}\end{aligned}$$

It follows that $\omega \to n\int_{t-1/n}^t f(s)\,ds$ is $\mathscr{F}_t$ measurable because $\mathscr{X}_{[0,t]}f$ is $\mathscr{F}_t \times \mathscr{B}([0,T])$ measurable by assumption that f is progressively measurable. ■

Observe that

$$\begin{aligned}\|f_n - f\|_{L^2(\Omega\times[0,T];\mathscr{L}_2(U,H))} &\equiv \left(\int_0^T \int_\Omega \|f_n - f\|^2_{\mathscr{L}_2} dP dt\right)^{1/2} \\ &= \left(\int_\Omega \int_0^T \left\|n\int_{-1/n}^0 (f(t+s) - f(t))\,ds\right\|^2_{\mathscr{L}_2} dt dP\right)^{1/2}\end{aligned}$$

From Minkowski's inequality,

$$\leq n\int_{-1/n}^0 \left(\int_\Omega \int_0^T \|f(t+s) - f(t)\|^2 dt dP\right)^{1/2} ds \tag{33.32}$$

so

$$\|f_n - f\|^2_{L^2(\Omega\times[0,T];\mathscr{L}_2(U,H))} \leq \int_\Omega n\int_{-1/n}^0 \int_0^T \|f(t+s) - f(t)\|^2 dt ds dP$$

Now

$$\begin{aligned} & n\int_{-1/n}^{0}\int_{0}^{T}\|f(t+s)-f(t)\|^2\,dtds \\ \leq\ & 2n\int_{-1/n}^{0}\int_{0}^{T}\|f(t+s)\|^2\,dtds+2n\int_{-1/n}^{0}\int_{0}^{T}\|f(t)\|^2\,dtds \\ \leq\ & 4\int_{0}^{T}\|f(t)\|^2\,ds \end{aligned}$$

which by definition is in $L^1(\Omega)$. Therefore, the integrands

$$n\int_{-1/n}^{0}\int_{0}^{T}\|f(t+s)-f(t)\|^2\,dtds$$

converge to 0 by continuity of translation in L^2 and are uniformly integrable. By Vitali convergence theorem, $\lim_{n\to\infty}\|f_n-f\|^2_{L^2(\Omega\times[0,T];\mathscr{L}_2(U,H))}=0$.

I want to use the fundamental inequality 33.31 which has only been presented above for f bounded. Therefore, let $g_n(t,\omega)\equiv P_{m_n}(f(t,\omega))$ where P_{m_n} is the projection onto $\overline{B(0,m_n)}$ in the Hilbert space $\mathscr{L}_2(U,H)$. As follows from the definition, one can obtain an inner product for the norm in this Banach space in the form $(f,g)\equiv\sum_{k=1}^{\infty}(f(e_i),g(e_i))_H$ where $\{e_i\}$ is some orthonormal basis for U. Now P_{m_n} is Lipschitz continuous and if m_n is large enough, $\|g_n-f_n\|_{L^2(\Omega\times[0,T];\mathscr{L}_2(U,H))}<2^{-n}$ and so $g_n(t,\omega)$ is bounded and continuous and adapted and

$$\lim_{n\to\infty}\|g_n-f\|_{L^2(\Omega\times[0,T];\mathscr{L}_2(U,H))}=0 \tag{33.33}$$

With this preparation, here is the main result.

Theorem 33.6.3 *Let f be adapted and in $L^2(\Omega\times[0,T];\mathscr{L}_2(U,H))$. Then there exists a sequence $\{g_n\}$ of adapted functions continuous in t such that*

$$\lim_{n\to\infty}\|g_n-f\|^2_{L^2(\Omega\times[0,T];\mathscr{L}_2(U,H))}=\lim_{n\to\infty}\int_\Omega\int_0^T\|g_n-f\|^2\,dtdP=0$$

Also it follows that $\int_0^t g_n dM$ is a Cauchy sequence in $\mathscr{M}_T^2(H)$ converging to a continuous martingale denoted as $\int_0^t f dM$ in $\mathscr{M}_T^2(H)$. In addition, for $f\in L^2(\Omega\times[0,T];\mathscr{L}_2(U,H))$, adapted, the fundamental inequality holds.

$$\frac{1}{2}E\left(\sup_{t\in[0,T]}\left\|\int_0^t f dM\right\|^2\right)\leq\int_\Omega\int_0^T\|f\|^2_{\mathscr{L}_2}\,dtdP$$

If M is only a local martingale, then the same inequality is valid. Also, if σ is any stopping time,

$$\int_0^{t\wedge\sigma}f dM=\int_0^t f dM^\sigma$$

Proof: It follows from the above argument there exists a sequence of adapted continuous, bounded, functions converging to f in $L^2(\Omega\times[0,T];\mathscr{L}_2(U,H))$. Therefore,

$$\lim_{m,n\to\infty}\frac{1}{2}E\left(\sup_{t\in[0,T]}\left\|\int_0^t(g_n-g_m)\,dM\right\|^2\right)\leq\lim_{m,n\to\infty}\int_\Omega\int_0^T\|g_n-g_m\|^2\,dtdP=0$$

and by completeness of $\mathscr{M}_T^2(H)$, $\int_0^t g_n dM$ converges to a continuous martingale which I can call $\int_0^t f dM$. It is clear that any two sequences give the same result from the inequality satisfied. Therefore, the stochastic integral $\int_0^t f dM$ is well defined. Also from the theory of $\mathscr{M}_T^2(H)$, there is a subsequence for which $\int_0^t g_n dM$ converges uniformly in t to $\int_0^t f dM$ off some set of measure zero.

In case M is only a local martingale, we see from approximating f with continuous in t and adapted and bounded functions g_n as above that the appropriate way to define $\int_0^t f dM$ is as $\int_0^{t\wedge\sigma_n} f dM \equiv \int_0^t f dM^{\sigma_n}$ where $\{\sigma_n\}$ is a localizing sequence for M.

The quadratic variation of M^{σ_n} is no more than the quadratic variation of M and so

$$\begin{aligned} & \lim_{m,n\to\infty} \frac{1}{2} E\left(\sup_{t\in[0,T]} \left\| \int_0^{t\wedge\sigma_n} (g_n - g_m)\, dM \right\|^2 \right) \\ \leq \quad & \lim_{m,n\to\infty} \int_\Omega \int_0^T \|g_n - g_m\|^2 \, dt dP = 0 \end{aligned}$$

Thus we can obtain $\int_0^t f dM^{\sigma_n}$ as a limit in $\mathscr{M}_T^2(H)$ as just done. Then one can define $\int_0^t f dM \equiv \lim_{n\to\infty} \int_0^t f dM^{\sigma_n}$ where σ_n is a localizing sequence for M. Also, we can pass to a limit as $n \to \infty$ using the monotone convergence theorem in the inequality

$$\frac{1}{2} E\left(\sup_{t\in[0,T]} \left\| \int_0^{t\wedge\sigma_n} f dM^{\sigma_n} \right\|^2 \right) \leq \int_\Omega \int_0^T \|f\|^2 \, dt dP$$

The stopping time has the effect of restricting the time interval, so as σ_n increases, one is taking sup over a larger set. That is why the monotone convergence theorem applies on the left side. Then

$$\frac{1}{2} E\left(\sup_{t\in[0,T]} \left\| \int_0^t f dM \right\|^2 \right) \leq \int_\Omega \int_0^T \|f\|^2 \, dt dP$$

As to the last claim about stopping times, it works for f bounded and continuous in t and adapted and M a martingale. Therefore, it also works for

$$f \in L^2(\Omega \times [0,T]; \mathscr{L}_2(U,H)).$$

In general, when M is only a local martingale with localizing sequence stopping times $\{\tau_n\}$, then M^σ is also a local martingale with localizing sequence $\{\tau_n\}$. I need to show that

$$\int_0^{t\wedge\sigma} f dM = \int_0^t f dM^\sigma$$

as local martingales. I need to show that

$$\int_0^{t\wedge\sigma\wedge\tau_n} f dM = \int_0^t f d(M^\sigma)^{\tau_n}$$

The equation is true because of the definition of $\int_0^{t\wedge\sigma} f dM$ in terms of the stopping times. Therefore, $\int_0^{t\wedge\sigma} f dM = \int_0^t f dM^\sigma$ as local martingales. Of course if M is a martingale, this is true also. ■

Bibliography

[1] **Apostol T. M.,** *Calculus Volume II Second edition,* Wiley *1969.*

[2] **Apostol, T. M.,** *Mathematical Analysis,* Addison Wesley Publishing Co., 1974.

[3] **Ash, Robert,** *Complex Variables*, Academic Press, 1971.

[4] **Baker, Roger**, *Linear Algebra*, Rinton Press 2001.

[5] **Balakrishnan A.V.,** *Applied Functional Analysis, Springer Verlag 1976.*

[6] **Billingsley P.,** *Probability and Measure,* Wiley, 1995.

[7] **Buck, R. C.** *Advanced Calculus* 2 edition. McGraw-Hill, 1965.

[8] **Brézis, H.** *Opérateurs maximaux monotones et semigroupes de contractions dans les espaces de Hilbert,* Math Studies, 5, North Holland, 1973.

[9] **Brézis, H.,** *Équations et inéquations non linéaires dans les espaces vectoriels en dualité ,* Ann. Inst. Fourier (Grenoble) 18 (1968) pp. 115-175.

[10] **Cheney, E. W.,** *Introduction To Approximation Theory,* McGraw Hill 1966.

[11] **Chow S.N. and Hale J.K.,** *Methods of bifurcation Theory ,* Springer Verlag, New York 1982.

[12] **Deimling K.** *Nonlinear Functional Analysis,* Springer-Verlag, 1985.

[13] **Diestal J.** and **Uhl J.,** *Vector Measures,* American Math. Society, Providence, R.I., 1977.

[14] **Dontchev A.L.** The Graves theorem Revisited, *Journal of Convex Analysis,* Vol. 3, 1996, No.1, 45-53.

[15] **Donal O'Regan, Yeol Je Cho, and Yu-Qing Chen,** Topological Degree Theory and Applications, Chapman and Hall/CRC 2006.

[16] **Dunford N.** and **Schwartz J.T.** *Linear Operators,* Interscience Publishers, a division of John Wiley and Sons, New York, part 1 1958, part 2 1963, part 3 1971.

[17] **Evans L.C. and Gariepy,** *Measure Theory and Fine Properties of Functions,* CRC Press, 1992.

[18] **Evans L.C.** *Partial Differential Equations,* Berkeley Mathematics Lecture Notes. 1993.

[19] **Fitzpatrick P. M.,** *Advanced Calculus a course in Mathematical Analysis,* PWS Publishing Company 1996.

[20] **Fonesca I. and Gangbo W.** *Degree theory in analysis and applications* Clarendon Press 1995.

[21] **Gasinski L.,Migorski S., and Ochal A.** Existence results for evolutionary inclusions and variational–hemivariational inequalities, *Applicable Analysis* 2014.

[22] **Gasinski L. and Papageorgiou N.,** *Nonlinear Analysis,* Volume 9, Chapman and Hall, 2006.

[23] **Gurtin M.** *An introduction to continuum mechanics,* Academic press 1981.

[24] **Gromes W.** Ein einfacher Beweis des Satzes von Borsuk. *Math. Z.* 178, pp. 399 -400 (1981).

[25] **Hardy, G.H., Littlewood, J.E. and Polya, G.,** Inequalities, Cambridge University Press 1964.

[26] **Hewitt E.** and **Stromberg K.** *Real and Abstract Analysis,* Springer-Verlag, New York, 1965.

[27] **Heinz, E.**An elementary analytic theory of the degree of mapping in n dimensional space. *J. Math. Mech. 8, 231-247* 1959

[28] **Hobson E.W.,** The Theory of functions of a Real Variable and the Theory of Fourier's Series V. 1, Dover 1957.

[29] **Hocking J. and Young G.,** Topology, Addison-Wesley Series in Mathematics, 1961.

[30] **Horn R. and Johnson C.,** *matrix Analysis,* Cambridge University Press, 1985.

[31] **Hu S. and Papageorgiou, N.** *Handbook of Multivalued Analysis*, Kluwer Academic Publishers (1997).

[32] **Karatzas and Shreve,** *Brownian Motion and Stochastic Calculus,* Springer Verlag, *1991.*

[33] **Kato T.** Perturbation Theory for Linear Operators, Springer, 1966.

[34] **Kreyszig E.** *Introductory Functional Analysis With applications,* Wiley 1978.

[35] **Kuratowski K.** and **Ryll-Nardzewski C.** A general theorem on selectors, *Bull. Acad. Pol. Sc.,* **13,** 397-403.

[36] **Kuttler K.L.,** *Modern Analysis* CRC Press 1998.

[37] **Kuttler K. L.**, *Basic Analysis*, Rinton

[38] **Kuttler K. L.**, *Linear Algebra and Analysis*, web page Web Page

[39] **Marsden J. E. and Hoffman J. M.,** *Elementary Classical Analysis,* Freeman, 1993.

[40] **McShane E. J.** *Integration,* Princeton University Press, Princeton, N.J. 1944.

[41] **Munkres, James R.,***Topology A First Course*, Prentice Hall, Englewood Cliffs, New Jersey 1975

[42] **Naylor A. and Sell R.,** *Linear Operator Theory in Engineering and Science*, Holt Rinehart and Winston, 1971.

[43] **Øksendal Bernt** *Stochastic Differential Equations*, Springer 2003.

[44] **Pettis, B.J.** On integration in vector spaces. Trans. Amer. Math.Soc. 44 277-304 (1938)

[45] **Prévôt C. and Röckner,** *A Concise Course on Stochastic Partial Differential Equations, Lecture notes in Mathematics, Springer 2007.*

[46] **Ray W.O.** *Real Analysis,* Prentice-Hall, 1988.

[47] **Rohatgi V. K.** *An Introduction to Probability Theory and Mathematical Statistics,* John Wiley and Sons, 1976.

[48] **Rudin W.,** *Principles of Mathematical Analysis*, McGraw Hill, 1976.

[49] **Rudin W.** *Real and Complex Analysis,* third edition, McGraw-Hill, 1987.

[50] **Rudin W.** *Functional Analysis,* second edition, McGraw-Hill, 1991.

[51] **Simon**, **J** *Compact sets in the space* $L^p(0,T;B)$*,* Ann. Mat. Pura. Appl. **146**(1987), 65-96.

[52] **Spanier E.**, *Algebraic Topology, McGraw Hill 1966.*

[53] **Spivak M.**, *Calculus On Manifolds*, Benjamin 1965.

[54] **Stromberg, K. R.** *Probability for analysts,* Chapman and Hall, *1994.*

[55] **Stroock D. W.** *Probability Theory An Analytic View,* Second edition*, Cambridge University Press, 2011*

[56] **Taylor A. E.** *General Theory of Functions and Integration,* Blaisdell Publishing, 1965

[57] **Vick, J**., *homology theory, An Introduction to Algebraic Topology, Academic Press 1973.*

[58] **Widder, D**. *Advanced Calculus,* second edition, Prentice Hall 1961.

[59] **Yosida K.** *Functional Analysis,* Springer-Verlag, New York, 1978.

Index

$(-\infty,\infty]$, 242
C^k, 197
C^1, 195
C^1 and differentiability, 195
C_c^∞, 374
C_c^m, 374
F_σ, 250
G_δ, 558
G_δ, 250
L^1
    approximation, 319
    weak compactness, 653
L^p
    definition, 367
    density of continuous functions, 371
    density of simple functions, 370
    density of smooth functions, 376
    norm, 367
    separability, 371
L^p
    reflexive, 653
    weak compactness, 653
L^1
    complex vector space, 291
$L^1(\Omega)$, 290
L^∞, 369
L^p
    compactness, 379
    continuity of translation, 373
$L^p(\Omega)$, 365
$L^p(\Omega;X)$, 695
L^1_{loc}, 374
$X \times Y$
    norm, 562
ε net, 78
$\mathbb{F}^n$, 99
$\mathscr{D}^*$, 455
$\mathscr{G}$, 383
$\mathscr{L}(X,Y)$, 559
    Banach space, 559
π systems, 247
σ algebra, 241

a-priori estimates, 169
absolutely continuous
    existence of derivative, 331
    function, 330
    function and measure, 330
    integral of derivative, 334
    integral of the derivative, 331
    Lipschitz, 331
    measure, 330
accumulation point, 71
adapted, 814, 845
adjoint linear map, 568
adjugate, 49
Alexander subbasis theorem, 524
algebra, 141
    Cartesian product, 546
    measure on algebra, 547
    recognizing one, 545
algebra of sets, 545
approximate identity, 374
apriori estimates, 169
arcwise connected, 90
    connected, 90, 144
area measure
    on manifold, 405
arithmetic mean, 237
Arzela Ascoli theorem
    Banach space, 716
Ascolil Arzela theorem
    general form, 87
at most countable, 61
axiom of choice, 57, 61
axiom of extension, 57
axiom of specification, 57
axiom of unions, 57

backwards Holder inequality, 381
backwards Minkowski inequality, 382
Baire
    category, 557, 558
Baire theorem, 122
Banach
    space, 557
Banach Alaoglu theorem, 580
Banach space, 106, 209, 367, 557, 739
Banach Steinhaus theorem, 560
barycenter, 159
basis, 101
basis of a topology, 517
Bernstein polynomial
    approximation of derivative, 132
Besicovitch
    covering theorem, 119, 321
Besicovitch covering theorem, 268
Bessel's inequality, 607, 641

Binet Cauchy formula, 45
block matrix, 15
block multiplication, 15, 17
Bochner integrable, 678
Borel
    measure, 262
Borel Cantelli lemma, 247, 743
Borel measure
    metric space, 262
    Polish space, 253
Borel regular, 459
Borel sets, 249
Borsuk, 429
Borsuk Ulam theorem, 432
Borsuk's theorem, 428
bounded continuous linear functions, 558
bounded linear maps, 209
    continuity, 126
bounded set, 110
Brouwer fixed point theorem, 344, 427, 600
    complact convex set, 165
Browder's lemma, 180, 282, 641
Burkholder Davis Gundy
    inequality, 905
Burkholder Davis Gundy inequality, 903

Cantor function, 273
Cantor set, 273
Caratheodory extension theorem, 547
Caratheodory functions, 278
Caratheodory's procedure, 260
Cariste fixed point theorem, 174
Cauchy Schwarz inequality, 104, 597
Cauchy sequence, 73
Cayley Hamilton theorem, 53
central limit theorem, 806
chain, 68
chain rule, 190
change of variables, 484
    linear map, 350
    linear maps, 352
    map not one to one, 358
change of variables general case, 358
characteristic function, 779
characteristic polynomial, 53
Clairaut's theorem, 201
Clarkson
    inequalities, 575
Clarkson inequalities, 576
Clarkson inequality
    $p \geq 2$, 572
    easy one, 573
closed disk, 110
closed graph theorem, 563
closed set, 72, 518
closed sets
    limit points, 72
closure of a set, 74, 75, 519
coarea formula, 480
cofactor, 47
cofactor identity, 204, 420
column rank, 50
compact map
    finite dimensional approximation, 165
compact set, 76, 520
compactness
    closed interval, 109
    equivalent conditions, 78
completely separable, 75
complex
    measure, 643
complex valued measurable functions, 289
components of a vector, 101
conditional expectation, 809
    Banach space, 728
    independence, 812
connected, 88
    open balls, 90, 144
connected component, 89
    boundary, 416
connected components, 89
    equivalence class, 89
    equivalence relation, 89
    open sets, 90
connected set
    continuous function, 91, 145
    continuous image, 88
connected sets
    intersection, 88
    intervals, 89
    real line, 89
    union, 88
continuity
    algebraic properties, 128
    bounded linear maps, 126
    coordinate maps, 128

uniform, 82
continuity of translation, 373
continuity set, 805
continuous function, 80, 519
maximum and minimum, 82
continuous functions
compact support, 296
equivalent conditions, 80
continuous image of compact set, 81
continuous martingale
not of bounded variation, 891
contraction map, 82
fixed point, 212
fixed point theorem, 82
convergence in measure, 315
convex, 739
set, 598
sets, 145
convex
functions, 379
convex combination, 120
convex function
continuous, 769
convex hull, 120, 157, 739
convolution, 374, 396
convolution of measures, 784
coordinate map, 111
coordinates, 157
countable, 61
covariance matrix, 787
covariation, 899
cowlicks, 432
Cramer's rule, 49
cylindrical set, 753, 778

definition of L^p, 367
definition of a C^k function, 197
density of $\mathscr{G}$ in L^p, 384
density of continuous functions in L^p, 371
derivative
chain rule, 190
continuity, 196
continuity of Gateaux derivative, 196
continuous, 189
continuous Gateaux derivatives, 193
Frechet, 188
Gateaux, 191, 193
generalized partial, 199
higher order, 196, 197
matrix, 191
partial, 199
second, 196
well defined, 189
derivative of inverse, 234
derivatives, 188
determinant
definition, 41
expansion along row, column, 47
matrix inverse, 48
permutation of rows, 42
product, 45
row, column operations, 43
symmetric definition, 43
transpose, 43
diameter of a set, 95
differentiable, 188
continuous, 189
continuous partials, 200
locally Lipschitz, 470
on relatively open set, 401
differentiable manifold, 412
differentiable map of Lebesgue measurable set, 353
differential equations
dependence on data, 586
global existence, 168
Peano existence theorem, 542
differentiation
Radon measures, 325
differentiation almost everywhere
monotone function, 271
dimension of a vector space, 102
Dini derivates, 270
directional derivative, 193
distance, 71
distance to a nonempty set, 241
distribution, 455, 743
distribution function, 303
divergence theorem, 409
dominated convergence
generalization, 295
dominated convergence theorem, 294, 697
Doob Dynkin lemma, 751
Doob estimate, 821
Doob's sub-martingale estimate, 854
dot product, 103

dual space, 567
duality maps, 594
dyadics, 126
Dynkin's lemma, 247
    alternative conditions, 274

Eberlein Smulian theorem, 584
Eggoroff theorem, 296
Egoroff theorem, 677
eigenvalue, 237
eigenvalues, 53
Ekeland variational principle, 172
elementary function, 911, 921
elementary functions
    linear space, 921
elementary matrices, 24
embedding into its double dual space, 569
equality of mixed partial derivatives, 203
equi-integrable, 299
equicontinuous, 86, 715
equivalence class, 63
equivalence relation, 63
events, 748
evolution equation
    continuous semigroup, 632
exchange theorem, 101
expectation, 750
    products of random variables, 750
exponential growth, 399
extending off closed set, 739
extention
    mapping, 739
extreme values theorem, 82

F sigma
    set, 250
Faddeyev theorem, 318
Fatou's lemma, 288
Fick's law, 413
filtration, 845
filtration
    normal, 845
finding the inverse, 37
finite intersection property, 520
finite measure
    regularity, 251, 745
first hitting time, 824
    closed set, 866
    open set, 866
fixed point property, 344, 446
fixed point theorem
    Cariste, 174
    Kakutani, 171
flip, 228
Fourier and inverse Fourier transforms, 387
Fourier series
    uniform convergence, 594
Fourier transform
    L^1, 389
    L^2, 391
    continuous, 389
    convolution, 390, 396
    in $\mathscr{G}^*$, 386
    of functions in G, 384
    convolution, 394
Fourier transform L^1, 389
Fourier transforms
    polynomial growth, 388
Frechet derivative, 188
Fredholm operator
    Banach space, 588
Fubini's theorem, 314
    Bochner integrable functions, 693
    general product measures, 312
function, 60
functions
    measurable, 241
fundamental theorem of calculus
    general Radon measures, 324
    Radon measures, 322

G delta, 250
Gamma function, 380, 464
gamma function, 207
Gateaux
    derivative, 449
Gateaux derivative, 191, 193
    continuous, 196
gauge function, 564
Gauss Jordan method for inverses, 37
generalized normal distribution, 790
geometric mean, 237
good lambda inequality, 305, 902
Gram Schmidt process, 109
Gram Schmidt process., 108
graph of a linear map, 562

Green's theorem, 513

Hahn
    decomposition, 307
Hahn Banach theorem, 565
    complex version, 566
Hahn decomposition, 667
Hahn Jordan decomposition, 667
Hamel basis, 121
Hardy's inequality, 380
Haursdorff measures, 457
Hausdorff
    maximal principle, 69
Hausdorff and Lebesgue measure, 464
Hausdorff dimension, 464
Hausdorff maximal principle, 524, 564
Hausdorff measure
    set of measure 0, 465
Hausdorff measures, 458
Hausdorff space, 518
Hermitian
    diagonalization, 20
    non-defective, 20
Hermitian matrix, 20
Hessian matrix, 223
higher order derivative
    multilinear form, 197
higher order derivatives, 196
    implicit function theorem, 216
    inverse function theorem, 216
Hilbert Schmidt
    operator, 614
Hilbert Schmidt theorem, 608, 689
Hilbert space, 597
hitting this before that, 881
Holder inequality
    backwards, 570
Holder's inequality, 106, 365
homeomorphism, 81

implicit function theorem, 212
    higher order derivatives, 216
increasing function
    existence of the derivative, 449
independent events, 748
independent random vectors, 749
independent sigma algebras, 749
indicator function
    approximation, 319
inner product space, 597
inner regular, 249, 744
    compact sets, 250
inner regularity, 263
Integral
    Riemann and Lebesgue, 554
integral
    continuous function, 137
    decreasing function, 283
    functions in L^1, 290
    linear, 290
integral over a measurable set, 295
integrals
    iterated, 140
integration
    with respect to a martingale, 911
integration by parts, 340
integration with respect to martingales
    Ito isometry, 916
interior point, 71
intermediate value theorem, 89
invariance of domain, 234, 430, 431
inverse, 30, 37
inverse
    left right, 38
    product of matrices, 30
    row reduced echelon form, 39
inverse function theorem, 215, 236
    higher order derivatives, 216
inverse image, 59
inverses and determinants, 48
invertible, 30
invertible maps, 209
    different spaces, 210
isodiametric inequality, 461, 463
isometric, 707
iterated integrals, 140
Ito isometry, 916

James map, 569
Jensen's inequality, 769
Jensens inequality, 379, 812
Jordan curve theorem, 436, 437
Jordan separation theorem, 439

Kakutani fixed point theorem, 171
Kantorovitch, 657

Kolmogorov Centsov theorem, 838, 841
Kolmogorov extension theorem, 550, 747
Kolmogorov zero one law, 756
Kolmogorov's inequality, 758
Kuratowski theorem, 278

Lagrange multipliers, 220, 221
Laplace expansion, 46
Laplace transform, 399, 543
least squares regression, 206
Lebesgue
    decomposition, 310
Lebesgue integral
    desires to be linear, 289
    nonnegative function, 284
    other definitions, 288
    simple function, 285
Lebesgue integral versus Riemann Stieltjes integral, 302
Lebesgue measurable function
    approximation with Borel measurable, 319
Lebesgue measure
    approximation with Borel sets, 319
    one dimensional, 264
    properties, 319
Lebesgue number, 77, 96
Lebesgue points, 322
Lebesgue Stieltjes measure, 263, 329
left inverse, 38, 39
Leray Schauder alternative, 167
Levy theorem, 800
lim inf, 66
    properties, 68
lim sup, 66
    properties, 68
lim sup and lim inf, 294
limit
    continuity, 187
    infinite limits, 185
    point, 71
limit of a function, 185
limit of a sequence, 72
    well defined, 72
limit point, 185, 518
limits
    combinations of functions, 185
    existence of limits, 66
    independent random variables, 757
limits and continuity, 187
Lindeloff property, 76
linear
    not continuous, 559
linear combination, 30, 44, 100
linear functional
    positive, 335
linear independence, 103
linear map of measurable set, 350
linear maps, 15
    closed, 563
    continuous, 558
    equivalent conditions, 558
linear relationship, 30
linear relationships
    row operations, 31
linear space, 99
linear transformation
    defined on a basis, 125
    dimension of vector space, 125
    rank m, 224
linear transformations
    a vector space, 125
    sum, 125
linearly dependent, 100
linearly independent, 100
linearly independent set
    enlarging to a basis, 103
Lipschitz
    continuous, 82
    functions, 449
    Gateaux derivatives, 450
Lipschitz function
    integral of its derivative, 449
Lipschitz functions, 401
    of measurable sets, 349
Lipschitz maps
    extension, 453
little o notation, 188
local martingale, 892
local maximum, 223
local minimum, 223
local submartingale, 892
localizing sequence, 892
locally compact, 520
locally compact , 520
locally finite, 377, 733

locally one to one, 238, 363
lower semicontinuous, 96
Lusin's theorem, 379
Lyapunov Schmidt procedure, 587

manifold
    boundary, 402
    orientable, 403
map
    C^1, 229
    primitive and flips, 229
martingale, 769
    quadratic variation, 894
martingales
    equiintegrable, 879
matrix
    left inverse, 49
    lower triangular, 49
    right inverse, 49
    right, left inverse, 49
    row, column, determinant rank, 50
    upper triangular, 49
matrix
    block multiplication, 17
    inverse, 30, 37
    more columns than rows, 35
    non zero kernel, 35
    partitioned, 17
    Schur's theorem, 19
matrix multiplication
    block, 17
maximal chain, 68
maximal estimate
    real sub-martingales, 829
maximal function
    Radon measures, 322
McShane's lemma, 659
mean value inequality, 192, 212
mean value theorem, 192, 212
    Cauchy, 137
measurability
    limit of simple functions, 244
measurable, 259
    complex valued, 289
    equivalent formulations, 242
    linear combinations, 289
    multifunction, 275
    multifunctions, 275
measurable complex functions
    simple functions, 293
measurable functions, 241
    approximation, 244
    pointwise limit, 242
    pointwise limits, 673
    simple functions, 242
measurable into $(-\infty, \infty]$, 242
measurable rectangles, 311
measurable representative, 703
measurable selection, 276
measurable sets, 259
measure, 246
    inner regular, 249
    on an algebra, 547
    outer regular, 249
    properties, 246
    vector, 643
measure space
    completion, 265
    regular, 370
measures
    complex, 643
    decreasing sequences of sets, 246
    increasing sequences of sets, 246
    regularity, 250, 253
    tight, 794
    weak convergence, 798
measures from outer measures, 260
metric, 71
    properties, 71
metric space, 71
    compact sets, 78
    complete, 74
    completely separable, 75
    open set, 71
    separable, 75
min max theorem, 147
min-max theorem, 147
Minkowski functional, 594
Minkowski inequality, 366
    integrals, 368
Minkowski inequality
    backwards, 570
Minkowski theorem
    for integrals, 381
Minkowski's inequality, 368
minor, 47

mixed partial derivatives, 201
modification, 835
mollifier, 374
monotone convergence theorem, 287
monotone function
    differentiable, 271
multi - index, 129
multi-index, 197, 383
multi-index notation, 197
multifunction
    strongly measurable, 277
multifunctions, 275
    measurability, 275
Muntz theorem, 154

negative part, 290
Neuman series, 209
Newton's method, 205
no retract onto boundary of ball, 427
non equal mixed partials
    example, 203
norm
    p norm, 106
normal, 411, 787, 801
normal filtration, 845
normal topological space, 519
Normed linear space, 106
nowhere differentiable functions, 592
nuclear operator, 612

one dimensional Stieltjes measure, 300
one point compactification, 522
open ball, 71
    open set, 71
open cover, 76, 520
open mapping theorem, 560
open set, 71
open sets, 517
    countable basis, 75
operator norm, 558
optional sampling theorem, 826
ordered
    partial, 68
    totally ordered, 68
oriented atlas, 403
orthonormal, 108
orthonormal set, 606
outer measure, 253
    measurable, 259
    total variation, 644
outer measure on real line, 256
outer regular, 249, 745
    G delta and F sigma sets, 251
outer regularity, 263

p norms, 107
paracompact space, 733
parallelogram identity, 641
partial derivatives, 191, 199
    continuous, 200
partial order, 68
partially ordered set, 68
partition
    one dimension, 333
partition of unity, 92, 528, 737
    infinitely differentiable, 375
    metric space, 736, 738
partitioned matrix, 17
permutation, 41
permutation matrix, 24
pi systems, 247
pivot column, 33
Plancherel theorem, 391, 392
point of density, 324
pointwise compact, 86
pointwise convergence, 85
polar coordinates
    integral, 360
polar decomposition, 647
Polish space, 76, 249
polynomial, 383
    in many variables, 129
polynomials, 129
positive, 617
positive linear functional, 319, 532
    measure, 335
positive part, 290
positive self adjoint
    products, 617
    roots, 618
power set, 57
precompact, 520, 542
predictable, 851
preserving distance, 22
primitive, 228
prior, 824

prior sets, 864
probability distribution function, 256
probability space, 743
product measure, 311, 314
    regular, 314
product of matrices
    inverse, 30
product space
    norm, 562
product topology, 519
products of Borel sets, 254, 339
progressively measurable, 845
progressively measurable
    integral of, 847
progressively measurable version, 849
projection in Hilbert space, 599
Prokhorov's theorem, 798

quadratic variation
    convergence in probability, 932
    fantastic properties, 915

Radon measure, 250, 267, 370, 371
Radon Nikodym
    Radon measures, 327
    theorem, 309
Radon Nikodym derivative, 309
Radon Nikodym property, 704
Radon Nikodym theorem
    Radon Measures, 327
random variable, 743
    distribution measure, 253
random vector, 743
    independent, 764
rank, 35
rank of a matrix, 50
rank theorem, 225
rational function, 129
real and imaginary parts, 289
recognizing a martingale
    stopping times, 887
refinement of a cover, 733
reflexive Banach Space, 570
reflexive Banach space, 653
    weak compactness, 584
regular, 249
regular measure space, 370
regular topological space, 518
regular values, 416
relative topology, 401
relatively open set, 401
resolvent, 627
retract, 344
    Banach space, 740
    closed and convex set, 740
    fixed point property, 344
reverse sub-martingale, 830
Reynolds
    transport formula, 478
Riemann integral, Lebesgue integral, 301
Riesz F., 382
Riesz map, 601
Riesz representation theorem, 709
    Hilbert space, 601
    locally compact Hausdorff space, 532
Riesz Representation theorem
    $C(X)$, 664
Riesz representation theorem L^p
    σ finite case, 652, 712
    finite measures, 648
Riesz representation theorem for L^1
    finite measures, 651
right inverse, 37–39
right polar factorization, 21
row equivalent, 33
row operations, 24
row operations
    linear relationships, 31
row rank, 50
row reduced echelon form, 33
row reduced echelon form
    existence, 33
    uniqueness, 34
Russell's paradox, 59

saddle point, 147
Sard's theorem, 353, 418
scalars, 99
Schaefer fixed point theorem, 167
Schauder fixed point
    approximate fixed point, 166
Schauder fixed point theorem, 167, 168
Schroder Bernstein theorem, 60
Schur's theorem, 19
Schwartz class, 394
second derivative, 196

second derivative test, 223, 224
sections of open sets, 199
self adjoint, 617
semigroup, 622
    adjoint, 636
    contraction
        bounded, 625
    generator, 624
    growth estimate, 625
    Hille Yosida theorem, 628
    strongly continuous, 624
separability of C(H), 797
separable metric space
    Lindeloff property, 76
separated sets, 88
separation theorem, 594
sequence, 72
    Cauchy, 73
    subsequence, 73
sequential weak* compactness, 582
sequentially compact set, 77
series
    double sum, 64
set
    F sigma, 250
set valued functions
    measurability, 275
sets, 57
    G delta, 250
sgn, 39
    uniqueness, 41
sigma algebra, 241
sigma finite
    measure space, 311
sign of a permutation, 41
signed measure, 307
    Hahn decomposition, 307
signed measures
    Hahn decomposition, 307
simple functions
    approximation, 242
singular values, 416
slicing measures, 537
smooth manifold, 402
Sobolev Space
    embedding theorem, 399
    equivalent norms, 399
Sobolev spaces, 399
space of continuous martingales, 884
    Hilbert space, 907
span, 44, 100
Sperner's lemma, 162
Steiner symetrization, 462
Stirling's formula, 348
stochastic integral
    as Stieltjes integral, 919
    elementary function, 921
    linear, 921
stochastic process, 814, 835
    descriptions, 835
Stone Weierstrass theorem, 526
Stone's theorem, 735
stong law of large numbers, 820
stopped martingale, 890, 892
stopped process, 881
stopped sub-martingale, 873
stopped submartingale, 892
stopping time, 823, 859, 863
stopping times
    conditional expectation, 825
strict convexity, 595
strictly convex
    norm, 147
strong law of large numbers, 761, 833
strongly measurable, 669
sub-martingale, 769, 814
sub-martingale convergence theorem, 817
    continuous case, 878
subbasis, 524
subspace, 100
sums
    independent random variables, 757
    independent variables, 759, 818
supermartingale, 814
support of a function, 92, 296, 528
symmetric derivative
    existence, 326
    measurable, 326
    Radon measure, 325
    upper and lower, 325
symmetric domain
    degree, 429
symmetric matrix, 20

tail event, 756
Taylor formula, 221

Taylor's formula, 222
Taylor's theorem, 222
Tietze extension theorem, 135
tight, 794
topological manifold, 401
    with boundary, 401
topological space, 517
total variation, 333, 644
totally bounded, 78, 86
totally ordered, 68
trace, 613
trajectories, 835
translation invariant, 264
triangle inequality, 106, 366
triangulated, 157
triangulation, 157
trivial, 100

uniform boundedness, 560
uniform boundedness theorem, 560
uniform continuity, 82
uniform continuity and compactness, 82
uniform convergence, 85
uniform convergence and continuity, 85
uniform convexity, 570, 595
uniformly bounded, 715
uniformly convex, 570
uniformly integrable, 297, 380
uniqueness of limits, 185
upcrossing, 770, 815
upper semicontinuity
    set valued map, 169
upper semicontinuous, 96
upper semicontinuous composition, 170
Urysohn's lemma, 521, 522

variance, 787
variational inequality, 599
vector
    measure, 643
vector measures, 703
vector space
    dimension, 102
vector space axioms, 99
vector valued function
    limit theorems, 185
vectors, 99
version, 835
Vitali
    convergence theorem, 297
Vitali
    convergence theorem, 697
Vitali convergence theorem, 300
Vitali cover, 268
Vitali covering theorem, 114
volume of unit ball, 464

weak and weak* topologies, 578
weak convergence, 595
    measures, 798
weak convergence of measures, 800
weak derivative, 456
weak topology, 579
weak$*$ measurable, 675
weak* topology, 579
    metric space, 581
weakly measurable, 669
wedge product, 493
    algebraic properties, 495
Weierstrass
    Stone Weierstrass theorem, 143
Weierstrass approximation
    estimate, 130
well ordered sets, 69

Young's inequality, 666