

Calculus of One and Many Variables

Kenneth Kuttler klkuttler@gmail.com

June 23, 2023

Contents

I	Functions of One Variable	15
1	Fundamental Concepts	17
1.1	Numbers and Simple Algebra	17
1.2	Exercises	20
1.3	Set Notation	21
1.4	Order	22
1.5	Exercises	27
1.6	The Binomial Theorem	27
1.7	Well Ordering Principle, Math Induction	28
1.8	Exercises	31
1.9	Completeness of \mathbb{R}	34
1.10	Existence of Roots	35
1.11	Completing the Square	37
1.12	Dividing Polynomials	38
1.13	The Complex Numbers	40
1.14	Polar Form of Complex Numbers	43
1.15	Roots of Complex Numbers	43
1.16	Exercises	45
1.17	Videos	48
2	Functions	49
2.1	General Considerations	49
2.2	Graphs of Functions and Relations	52
2.3	Circular Functions	53
2.3.1	Reference Angles and Other Identities	60
2.3.2	The $\sin(x)/x$ Inequality	62
2.3.3	The Area of a Circular Sector	65
2.4	Exercises	67
2.5	Exponential and Logarithmic Functions	71
2.6	The Function b^x	75
2.7	Applications	77
2.7.1	Interest Compounded Continuously	77
2.7.2	Exponential Growth and Decay	77
2.7.3	The Logistic Equation	79
2.8	Using MATLAB to Graph	80
2.9	Exercises	80

2.10	Videos	82
3	Sequences and Compactness	83
3.1	Sequences	83
3.2	Exercises	84
3.3	The Limit of a Sequence	85
3.4	The Nested Interval Lemma	92
3.5	Exercises	92
3.6	Compactness	94
3.6.1	Sequential Compactness	94
3.6.2	Closed and Open Sets	95
3.7	Cauchy Sequences	98
3.8	Exercises	98
3.9	Videos	101
4	Continuous Functions and Limits of Functions	103
4.1	An Equivalent Formulation of Continuity	107
4.2	Exercises	107
4.3	The Extreme Values Theorem	109
4.4	The Intermediate Value Theorem	110
4.5	Continuity of the Inverse	112
4.6	Exercises	112
4.7	Uniform Continuity	113
4.8	Examples of Continuous Functions	114
4.9	Sequences of Functions	115
4.10	Polynomials and Continuous Functions	117
4.11	Exercises	120
4.12	Limit of a Function	121
4.13	Exercises	125
4.14	Videos	127
5	The Derivative	129
5.1	The Definition of the Derivative	129
5.2	Finding the Derivative	132
5.3	Derivatives of Inverse Functions	133
5.4	Circular Functions and Inverses	135
5.5	Exponential Functions and Logarithms	137
5.6	The Complex Exponential	138
5.7	Related Rates and Implicit Differentiation	139
5.8	Exercises	140
5.9	Local Extreme Points	141
5.10	Exercises	143
5.11	Mean Value Theorem	147
5.12	Exercises	148
5.13	First and Second Derivative Tests	150
5.14	Exercises	151
5.15	Taylor Series Approximations	152
5.16	Exercises	154

5.17	L'Hôpital's Rule	155
5.18	Interest Compounded Continuously	159
5.19	Exercises	160
5.20	Videos	162
6	Infinite Series	163
6.1	Basic Considerations	163
6.2	Absolute Convergence	166
6.3	Ratio and Root Tests	169
6.4	Exercises	171
6.5	Convergence Because of Cancellation	172
6.6	Double Series	173
6.7	Exercises	176
6.8	Series of Functions	178
6.9	Exercises	179
7	The Integral	181
7.1	The Definition of the Integral from Antiderivatives	183
7.2	Uniform Convergence and the Integral	188
7.3	The Riemann Darboux Integral*	188
7.4	Exercises	195
7.5	Videos	200
8	Methods for Finding Antiderivatives	201
8.1	The Method of Substitution	201
8.2	Exercises	203
8.3	Integration by Parts	205
8.4	Exercises	207
8.5	Trig. Substitutions	209
8.6	Exercises	213
8.7	Partial Fractions	214
8.8	Rational Functions of Trig. Functions	220
8.9	Using MATLAB	221
8.10	Exercises	222
8.11	Videos	224
9	A Few Standard Applications	225
9.1	Lengths of Curves and Areas of Surfaces of Revolution	226
9.1.1	Lengths	226
9.1.2	Surfaces of Revolution	228
9.2	Exercises	230
9.3	Force on a Dam and Work	232
9.3.1	Force on a Dam	232
9.3.2	Work	232
9.4	Using MATLAB	234
9.5	Exercises	234

10 Improper Integrals and Stirling's Formula	237
10.1 Stirling's Formula	237
10.2 The Gamma Function	240
10.3 Laplace Transforms	243
10.4 Exercises	246
11 Power Series	253
11.1 Functions Defined in Terms of Series	253
11.2 Operations on Power Series	255
11.3 Power Series for Some Known Functions	258
11.4 The Binomial Theorem	258
11.5 Exercises	260
11.6 Multiplication of Power Series	262
11.7 Exercises	263
11.8 Some Other Theorems	265
11.9 Some Historical Observations	269
12 Polar Coordinates	271
12.1 Graphs in Polar Coordinates	272
12.2 The Area in Polar Coordinates	273
12.3 The Acceleration in Polar Coordinates	275
12.4 The Fundamental Theorem of Algebra	277
12.5 Polar Graphing in MATLAB	278
12.6 Exercises	279
13 Algebra and Geometry of \mathbb{R}^p	281
13.1 \mathbb{R}^p	281
13.2 Algebra in \mathbb{R}^p	283
13.3 Geometric Meaning Of Vector Addition In \mathbb{R}^3	284
13.4 Lines	285
13.5 Distance in \mathbb{R}^p	288
13.6 Geometric Meaning of Scalar Multiplication in \mathbb{R}^3	291
13.7 Exercises	292
14 Vector Products	295
14.1 The Dot Product	295
14.2 Geometric Significance of the Dot Product	297
14.2.1 The Angle Between Two Vectors	297
14.2.2 Work and Projections	299
14.3 Exercises	302
14.4 The Cross Product	303
14.4.1 The Box Product	306
14.5 Proof of the Distributive Law	307
14.5.1 Torque	308
14.5.2 Center of Mass	309
14.5.3 Angular Velocity	310
14.6 Vector Identities and Notation	311
14.7 Planes	314
14.8 Exercises	317

15 Sequences, Compactness, and Continuity	321
15.1 Sequences of Vectors	321
15.2 Open and Closed Sets	322
15.3 Cartesian Products	324
15.4 Sequential Compactness	325
15.5 Vector Valued Functions	326
15.6 Continuous Functions	327
15.7 Sufficient Conditions for Continuity	328
15.8 Limits of a Function of Many Variables	329
15.9 Vector Fields	332
15.10 MATLAB and Vector Fields	333
15.11 Exercises	333
15.12 Extreme Value Theorem, Uniform Continuity	335
15.13 Convergence of Functions	336
15.14 Fundamental Theorem of Algebra	337
15.15 Exercises	338
16 Space Curves	343
16.1 Using MATLAB to Graph Space Curves	343
16.2 The Derivative and Integral	344
16.2.1 Geometric and Physical Significance of the Derivative	345
16.2.2 Differentiation Rules	347
16.2.3 Leibniz's Notation	349
16.3 Arc Length and Orientations	350
16.4 Arc Length and Parametrizations*	353
16.4.1 Hard Calculus	353
16.4.2 Independence of Parametrization	355
16.5 Exercises	356
16.6 Motion on Space Curves	358
16.6.1 Some Simple Techniques	360
16.7 Geometry of Space Curves*	362
16.8 Exercises	364
17 Some Physical Applications	367
17.1 Spherical and Cylindrical Coordinates	367
17.2 Exercises	369
17.3 Planetary Motion	370
17.3.1 The Equal Area Rule, Kepler's Second Law	371
17.3.2 Inverse Square Law, Kepler's First Law	372
17.3.3 Kepler's Third Law	374
17.4 The Angular Velocity Vector	375
17.5 Angular Velocity Vector on Earth	377
17.6 Coriolis Force and Centripetal Force	378
17.7 Coriolis Force on the Rotating Earth	379
17.8 The Foucault Pendulum*	381
17.9 Exercises	383

II Functions of Many Variables	387
18 Linear Functions	389
18.1 The Matrix of a Linear Transformation	389
18.2 Row Operations and Linear Equations	396
18.2.1 Using MATLAB	405
18.2.2 Uniqueness	405
18.2.3 The Inverse	406
18.2.4 MATLAB and Matrix Arithmetic	408
18.3 Exercises	409
18.4 Subspaces Spans and Bases	414
18.5 Linear Independence	416
18.6 Exercises	419
19 Eigenvalues and Eigenvectors	425
19.1 Definition of Eigenvalues	425
19.2 An Introduction to Determinants	426
19.2.1 Cofactors and 2×2 Determinants	426
19.2.2 The Determinant of a Triangular Matrix	428
19.2.3 Properties of Determinants	429
19.2.4 Finding Determinants Using Row Operations	431
19.3 MATLAB and Determinants	433
19.4 Applications	433
19.4.1 A Formula for the Inverse	433
19.4.2 Finding Eigenvalues Using Determinants	435
19.5 MATLAB and Eigenvalues	436
19.6 Matrices and the Dot Product	436
19.7 Distance and Orthogonal Matrices	437
19.8 Diagonalization of Symmetric Matrices	438
19.9 Exercises	442
20 The Mathematical Theory of Determinants*	449
20.1 The Function sgn	449
20.2 The Determinant	451
20.2.1 The Definition	451
20.2.2 Permuting Rows Or Columns	452
20.2.3 A Symmetric Definition	453
20.2.4 The Alternating Property of the Determinant	453
20.2.5 Linear Combinations and Determinants	454
20.2.6 The Determinant of a Product	455
20.2.7 Cofactor Expansions	456
20.2.8 Row, Column, and Determinant Rank	457
20.2.9 Formula for the Inverse	459
20.2.10 The Cayley Hamilton Theorem	460
20.2.11 Cramer's Rule	462
20.3 p Dimensional Parallelepipeds	462

21 Functions of Many Variables	465
21.1 Graphs	465
21.2 Review of Limits	465
21.3 Exercises	467
21.4 Directional and Partial Derivatives	468
21.4.1 The Directional Derivative	468
21.4.2 Partial Derivatives	469
21.5 Exercises	471
21.6 Mixed Partial Derivatives	473
21.7 Partial Differential Equations	474
21.8 Exercises	475
22 Derivative of a Functions of Many Variables	477
22.1 The Derivative of Functions of One Variable	477
22.2 The Derivative	479
22.3 Exercises	484
22.4 C^1 Functions	486
22.5 The Chain Rule	489
22.5.1 The Chain Rule for Functions of One Variable	489
22.5.2 The Chain Rule for Functions of Many Variables	489
22.6 Exercises	495
22.6.1 Related Rates Problems	496
22.6.2 The Derivative of the Inverse Function	497
22.7 Exercises	498
22.8 The Gradient	500
22.9 The Gradient and Tangent Planes	502
22.10 Exercises	503
23 Optimization	505
23.1 Local Extrema	505
23.2 Exercises	507
23.3 The Second Derivative Test	508
23.4 Exercises	511
23.5 Lagrange Multipliers	513
23.6 Exercises	518
23.7 Proof of the Second Derivative Test*	522
24 Implicit Function Theorem*	525
24.1 More Continuous Partial Derivatives	530
24.2 The Method of Lagrange Multipliers	531
25 Line Integrals	533
25.1 Line Integrals and Work	533
25.2 Conservative Fields and Notation	536
25.3 Exercises	537

26 The Riemannn Integral on \mathbb{R}^p	539
26.1 Methods for Double Integrals	539
26.1.1 Density and Mass	544
26.2 Exercises	544
26.3 Methods for Triple Integrals	546
26.3.1 Definition of the Integral	546
26.3.2 Iterated Integrals	546
26.4 Exercises	549
26.4.1 Mass and Density	551
26.5 Exercises	552
27 The Integral in Other Coordinates	555
27.1 Polar Coordinates	555
27.2 Exercises	557
27.3 Cylindrical and Spherical Coordinates	558
27.3.1 Volume and Integrals in Cylindrical Coordinates	559
27.3.2 Volume and Integrals in Spherical Coordinates	560
27.4 Exercises	567
27.5 The General Procedure	568
27.6 Exercises	571
27.7 The Moment of Inertia and Center of Mass	573
27.8 Exercises	574
28 The Integral on Two Dimensional Surfaces in \mathbb{R}^3	577
28.1 The Two Dimensional Area in \mathbb{R}^3	577
28.2 Surfaces of the Form $z = f(x, y)$	581
28.3 MATLAB and Graphing Surfaces	583
28.4 Piecewise Defined Surfaces	583
28.5 Flux Integrals	584
28.6 Exercises	584
29 Calculus of Vector Fields	587
29.1 Divergence and Curl of a Vector Field	587
29.1.1 Vector Identities	588
29.1.2 Vector Potentials	589
29.1.3 The Weak Maximum Principle	590
29.2 Exercises	591
29.3 The Divergence Theorem	592
29.3.1 Coordinate Free Concept of Divergence	596
29.4 Applications of the Divergence Theorem	597
29.4.1 Hydrostatic Pressure	597
29.4.2 Archimedes Law of Buoyancy	598
29.4.3 Equations of Heat and Diffusion	598
29.4.4 Balance of Mass	600
29.4.5 Balance of Momentum	600
29.4.6 Frame Indifference	606
29.4.7 Bernoulli's Principle	607
29.4.8 The Wave Equation	608

29.4.9 A Negative Observation	608
29.4.10 Volumes of Balls in \mathbb{R}^n	608
29.4.11 Electrostatics	610
29.5 Exercises	612
30 Stokes and Green's Theorems	615
30.1 Green's Theorem	615
30.2 Exercises	617
30.3 Stoke's Theorem from Green's Theorem	619
30.3.1 The Normal and the Orientation	621
30.3.2 The Mobeus Band	623
30.4 A General Green's Theorem	624
30.5 Conservative Vector Fields	625
30.5.1 Some Terminology	627
30.6 Exercises	628
31 Curvilinear Coordinates	633
31.1 Basis Vectors	633
31.2 Exercises	636
31.3 Curvilinear Coordinates	637
31.4 Exercises	640
31.5 Transformation of Coordinates.	642
31.6 Differentiation and Christoffel Symbols	643
31.7 Gradients and Divergence	645
31.8 Exercises	648
32 Measures and Integrals	651
32.1 Countable Sets	652
32.2 Simple Functions, σ Algebras, Measurability	654
32.3 Measures and Outer Measures	660
32.4 Measures from Outer Measures	661
32.5 Riemann Integrals for Decreasing Functions	665
32.6 Lebesgue Integrals of Nonnegative Functions	666
32.7 Nonnegative Simple Functions	667
32.8 The Monotone Convergence Theorem	669
32.9 The Integral's Righteous Algebraic Desires	670
32.10 Integrals of Real Valued Functions	670
32.11 Dynkin's Lemma	673
32.12 Product Measures	675
32.13 Exercises	677
33 The Lebesgue Measure and Integral in \mathbb{R}^p	681
33.1 An Outer Measure on $\mathcal{P}(\mathbb{R})$	681
33.2 One Dimensional Lebesgue Measure	682
33.3 The Lebesgue Integral and Riemann Integral	683
33.4 p Dimensional Lebesgue Measure and Integrals	684
33.4.1 Iterated Integrals	684
33.4.2 p Dimensional Lebesgue Measure and Integrals	685
33.5 Lebesgue Measure and Linear Maps	686

33.6 Change of Variables for Nonlinear Maps	688
33.7 Exercises	691

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. klkuttler@gmail.com I do constantly upgrade this book when I find things which could be improved.

Introduction

This book is a discussion of calculus of functions of real variables. It is written to be a first course in Calculus and to be used in the first three semesters of calculus. However, it would also work as an advanced calculus book for those who have had the typical undergraduate calculus course by emphasizing those chapters which are more theoretical in nature, although there is a lot more on one variable ideas in my single variable advanced calculus text, *Analysis of Functions of One Variable*. I am assuming the reader has had college algebra and trigonometry although most of what is needed is reviewed. This book is an extensively re-written version of my earlier Calculus text published with World Scientific [21]. This one has a lot more on algebra and fundamental ideas. It is more theoretical than my earlier book and somewhat shorter. It is not as close to the voluminous standard texts on Calculus. I have also included in the text simple techniques for using MATLAB which I think will be very helpful. It does include all the standard techniques however which I have attempted to present as simply as possible. I believe, based on my experience teaching engineering math, that these techniques are usually not mastered by students in their introductory calculus course, especially the technique of partial fractions. I hope that my presentation will be short enough to be easily remembered.

That which has the most mathematical significance is often marginalized, thus ignoring what was learned early in the nineteenth century. In general, existence theorems are neglected, which results in incomplete explanations of many of the most important theorems like mean value theorem, fundamental theorem of calculus etc. Calculus is not geometry. Neither is it algebra, and to neglect that which is tied to this observation is to misrepresent what the subject is all about. However, the book will work for a course in which these important topics are left for interested students.

There are proofs of the intermediate value theorem which is due to Bolzano and dates from around 1817 and the extreme value theorem, also done by Bolzano in the 1830's and later by Weierstrass. I will also show why the integral of a continuous function exists in two different ways, one quite unusual. The integral was of interest throughout the nineteenth century, beginning with the work of Cauchy. I don't understand how anyone can make sense of later courses like differential equations without this. What good is Picard iteration if you don't even know why the integrals you are writing down exist?

Also I am trying to present all of the main ideas in a somewhat shorter book than usual. I don't understand why it should be necessary to take over 1000 pages, even with the inclusion of physical applications, which were the motivation for developing calculus in the first place. I hope I have enough exercises but it is also not clear to me why such long lists of mostly routine or technical exercises are needed. This book is not like the "Think and do" books I had in elementary school in the 1950's. The book itself comes to a little more than 700 pages.

I have introduced some of the most important ideas more than once. For example, existence of roots illustrates in a specific case the intermediate value theorem presented in full generality later. The ideas leading to the integral are first encountered early in the book in a discussion of the logarithm. This is an “early transcendentals” book. Of course it is more efficient to present these functions only once later on, but I believe that calculus is about continuous functions, integrals, completeness of \mathbb{R} , and derivatives and that the ideas associated with these things should be emphasized.

The book is divided into two parts. The first part is on functions of one variable with some important theory pertaining to the second part (compact sets, extreme values theorem, etc.). There is some repetition here since it is done first for functions of one variable. The second part is on vector valued functions of many variables and is devoted to the standard topics in vector calculus. I have in mind the first eight or nine chapters for the first semester of calculus and the next eight for the second. Then the third semester would consist of whatever can be covered in the remainder of the book. There is more there than can be included in one semester.

The reason for the chapters on linear algebra is that multi-variable calculus is mostly based on reduction to linear algebra ideas. Contrary to the pretensions of virtually all standard texts, there is such a thing as the derivative of a function of many variables, it is very important, and it is a linear transformation. This seems to be the best kept secret in undergraduate math. I think multi-variable calculus would be better understood after a course on linear algebra. After all, linear functions are easier than nonlinear ones. Shouldn't we study the easy case first? If this is done, the chapters on linear algebra can be omitted or used as a review. On the other hand, the more significant course in undergraduate math is linear algebra, not calculus. Thus, if linear algebra is to come after multi-variable calculus, these chapters will help make the linear algebra course easier to master and make it possible to offer a better linear algebra course, since the stuff involving row operations and eigenvalues will have been seen already in calculus. Either way, exposure to a limited amount of linear algebra is a good idea in a multi-variable calculus book.

There is more in the book than will typically be discussed. Chapter 17 for example, is not usually included in beginning calculus but gives physical applications which illustrate the use of calculus and vector methods. To begin with, there are a few prerequisite topics. These can be referred to as needed.

At the end of many chapters and possibly at other places there are links to on line explanations. I am presently working on these.

Part I

Functions of One Variable

Chapter 1

Fundamental Concepts

1.1 Numbers and Simple Algebra

To begin with, consider the real numbers, denoted by \mathbb{R} , as a line extending infinitely far in both directions. In this book, the notation, \equiv indicates something is being defined. Thus the integers are defined as

$$\mathbb{Z} \equiv \{\dots -1, 0, 1, \dots\},$$

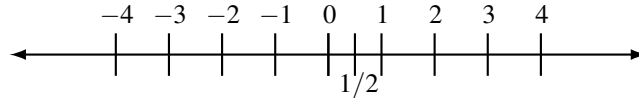
the natural numbers,

$$\mathbb{N} \equiv \{1, 2, \dots\}$$

and the rational numbers, defined as the numbers which are the quotient of two integers.

$$\mathbb{Q} \equiv \left\{ \frac{m}{n} \text{ such that } m, n \in \mathbb{Z}, n \neq 0 \right\}$$

are each subsets of \mathbb{R} as indicated in the following picture.



As shown in the picture, $\frac{1}{2}$ is half way between the number 0 and the number, 1. By analogy, you can see where to place all the other rational numbers. It is assumed that \mathbb{R} has the following algebra properties, listed here as a collection of assertions called axioms. These properties will not be proved which is why they are called axioms rather than theorems. In general, axioms are statements which are regarded as true. Often these are things which are “self evident” either from experience or from some sort of intuition but this does not have to be the case. In the following list, it is always assumed that $0 \neq 1$ since otherwise, everything reduces to consideration of 0. You would have $x = 1x = 0x = 0$ so all numbers would be 0. We are not interested in this.

Axiom 1.1.1 $x + y = y + x$, (*commutative law for addition*)

Axiom 1.1.2 $x + 0 = x$, (*additive identity*).

Axiom 1.1.3 For each $x \in \mathbb{R}$, there exists $-x \in \mathbb{R}$ such that $x + (-x) = 0$, (*existence of additive inverse*).

Axiom 1.1.4 $(x + y) + z = x + (y + z)$, (associative law for addition).

Axiom 1.1.5 $xy = yx$, (commutative law for multiplication).

Axiom 1.1.6 $(xy)z = x(yz)$, (associative law for multiplication).

Axiom 1.1.7 $1x = x$, (multiplicative identity).

Axiom 1.1.8 For each $x \neq 0$, there exists x^{-1} such that $xx^{-1} = 1$. (existence of multiplicative inverse).

Axiom 1.1.9 $x(y + z) = xy + xz$. (distributive law).

These axioms are known as the field axioms and any set (there are many others besides \mathbb{R}) which has two such operations satisfying the above axioms is called a field. Division and subtraction are defined in the usual way by $x - y \equiv x + (-y)$ and $x/y \equiv x(y^{-1})$. It is assumed that the reader is completely familiar with these axioms in the sense that he or she can do the usual algebraic manipulations taught in high school and junior high algebra courses. The axioms listed above are just a careful statement of exactly what is necessary to make the usual algebraic manipulations valid.

A word of advice regarding division and subtraction is in order here. Whenever you feel a little confused about an algebraic expression which involves division or subtraction, think of division as multiplication by the multiplicative inverse as just indicated and think of subtraction as addition of the additive inverse. Thus, when you see x/y , think $x(y^{-1})$ and when you see $x - y$, think $x + (-y)$. In many cases the source of confusion will disappear almost magically.

The reason for this is that subtraction and division do not satisfy the associative law. This means there is a natural ambiguity in an expression like $6 - 3 - 4$. Do you mean $(6 - 3) - 4 = -1$ or $6 - (3 - 4) = 6 - (-1) = 7$? It makes a difference doesn't it? However, the so called binary operations of addition and multiplication are associative and so no such confusion will occur. It is conventional to simply do the operations in order of appearance reading from left to right. Thus, if you see $6 - 3 - 4$, you would normally interpret it as the first of the above alternatives, but what if you grew up reading Hebrew or Arabic which reads from right to left according to my understanding? Shouldn't mathematics be independent of such things? Subtraction and division are abominations.

In the first part of the following theorem, the claim is made that the additive inverse and the multiplicative inverse are unique. This means that for a given number, only one number has the property that it is an additive inverse and that, given a nonzero number, only one number has the property that it is a multiplicative inverse. The significance of this is that if you are wondering if a given number is the additive inverse of a given number, all you have to do is to check and see if it acts like one.

Theorem 1.1.10 *The above axioms imply the following.*

1. *The multiplicative inverse and additive inverse are unique.*
2. $0x = 0$, $-(-x) = x$,
3. $(-1)(-1) = 1$, $(-1)x = -x$
4. *If $xy = 0$ then either $x = 0$ or $y = 0$.*

Proof: Suppose then that x is a real number and that $x + y = 0 = x + z$. It is necessary to verify $y = z$. From the above axioms, there exists an additive inverse, $-x$ for x . Therefore,

$$-x + 0 = (-x) + (x + y) = (-x) + (x + z)$$

and so by the associative law for addition, $((-x) + x) + y = ((-x) + x) + z$ which implies $0 + y = 0 + z$. Now by the definition of the additive identity, this implies $y = z$. You should prove the multiplicative inverse is unique.

Consider 2. It is desired to verify $0x = 0$. From the definition of the additive identity and the distributive law it follows that

$$0x = (0 + 0)x = 0x + 0x.$$

From the existence of the additive inverse and the associative law it follows

$$\begin{aligned} 0 &= (-0x) + 0x = (-0x) + (0x + 0x) \\ &= ((-0x) + 0x) + 0x = 0 + 0x = 0x \end{aligned}$$

To verify the second claim in 2., it suffices to show x acts like the additive inverse of $-x$ in order to conclude that $-(-x) = x$. This is because it has just been shown that additive inverses are unique. By the definition of additive inverse, $x + (-x) = 0$ and so $x = -(-x)$ as claimed.

To demonstrate 3., $(-1) + (-1)(-1) = (-1)(1 + (-1)) = (-1)0 = 0$. It follows from 1. and 2. that $1 = -(-1) = (-1)(-1)$. To verify $(-1)x = -x$, use 2. and the distributive law to write

$$x + (-1)x = x(1 + (-1)) = x0 = 0.$$

Therefore, by the uniqueness of the additive inverse proved in 1., it follows $(-1)x = -x$ as claimed.

To verify 4., suppose $x \neq 0$. Then x^{-1} exists by the axiom about the existence of multiplicative inverses. Therefore, by 2. and the associative law for multiplication,

$$y = (x^{-1}x)y = x^{-1}(xy) = x^{-1}0 = 0.$$

This proves 4. ■

Recall the notion of something raised to an integer power. Thus $y^2 = y \times y$ and $b^{-3} = \frac{1}{b^3}$ etc.

Also, there are a few **conventions** related to the order in which operations are performed. Exponents are always done before multiplication. Thus $xy^2 = x(y^2)$ and is not equal to $(xy)^2$. Division or multiplication is always done before addition or subtraction. Thus $x - y(z + w) = x - [y(z + w)]$ and is not equal to $(x - y)(z + w)$. Parentheses are done before anything else. Be very careful of such things since they are a source of mistakes. When you have doubts, insert parentheses to describe exactly what is meant.

Also recall summation notation.

Definition 1.1.11 Let x_1, x_2, \dots, x_m be numbers. Then

$$\sum_{j=1}^m x_j \equiv x_1 + x_2 + \dots + x_m.$$

Thus this symbol, $\sum_{j=1}^m x_j$ means to take all the numbers, x_1, x_2, \dots, x_m and add them. Note the use of the j as a generic variable which takes values from 1 up to m . This notation will be used whenever there are things which can be added, not just numbers. The notation $\sum_{i \in S} x_i$ means to consider all x_i for $i \in S$ and add them.

Also, $\prod_{i=1}^m x_i$ means to multiply all the x_i together: $\prod_{i=1}^m x_i \equiv x_1 x_2 \cdots x_m$

As an example of the use of this notation, you should verify the following.

Example 1.1.12 $\sum_{k=1}^6 (2k+1) = 48$, $\prod_{i=1}^3 (i+1) = 24$.

Be sure you understand why $\sum_{k=1}^{m+1} x_k = \sum_{k=1}^m x_k + x_{m+1}$. As a slight generalization of this notation, $\sum_{j=k}^m x_j \equiv x_k + \cdots + x_m$. It is also possible to change the variable of summation. $\sum_{j=1}^m x_j = x_1 + x_2 + \cdots + x_m$ while if r is an integer, the notation requires $\sum_{j=1+r}^{m+r} x_{j-r} = x_1 + x_2 + \cdots + x_m$ and so $\sum_{j=1}^m x_j = \sum_{j=1+r}^{m+r} x_{j-r}$.

Summation notation will be used throughout the book whenever it is convenient to do so.

When you have algebraic expressions, you treat the variables like they are numbers and add like you would normally do. For example, consider the following.

Example 1.1.13 Add the fractions, $\frac{x}{x^2+y} + \frac{y}{x-1}$.

You add these just like fractions. Write the first expression as $\frac{x(x-1)}{(x^2+y)(x-1)}$ and the second as $\frac{y(x^2+y)}{(x-1)(x^2+y)}$. Then since these have the same common denominator, you add them as follows.

$$\frac{x}{x^2+y} + \frac{y}{x-1} = \frac{x(x-1)}{(x^2+y)(x-1)} + \frac{y(x^2+y)}{(x-1)(x^2+y)} = \frac{x^2 - x + yx^2 + y^2}{(x^2+y)(x-1)}.$$

I assume the reader knows all about this kind of thing.

1.2 Exercises

1. Consider the expression $x + y(x+y) - x(y-x) \equiv f(x,y)$. Find $f(-1,2)$.
2. Show $-(ab) = (-a)b$.
3. Show on the number line the effect of multiplying a number by -1 .
4. Add the fractions $\frac{x}{x^2-1} + \frac{x-1}{x+1}$.
5. Find a formula for $(x+y)^2$, $(x+y)^3$, and $(x+y)^4$. Based on what you observe for these, give a formula for $(x+y)^8$.
6. When is it true that $(x+y)^n = x^n + y^n$?
7. Find the error in the following argument. Let $x = y = 1$. Then $xy = y^2$ and so $xy - x^2 = y^2 - x^2$. Therefore, $x(y-x) = (y-x)(y+x)$. Dividing both sides by $(y-x)$ yields $x = x+y$. Now substituting in what these variables equal yields $1 = 1+1$.

8. Find the error in the following argument. $\sqrt{x^2+1} = x+1$ and so letting $x = 2$, $\sqrt{5} = 3$. Therefore, $5 = 9$.
9. Find the error in the following. Let $x = 1$ and $y = 2$. Then $\frac{1}{3} = \frac{1}{x+y} = \frac{1}{x} + \frac{1}{y} = 1 + \frac{1}{2} = \frac{3}{2}$. Then cross multiplying, yields $2 = 9$.
10. Find the error in the following argument. Let $x = 3$ and $y = 1$. Then $1 = 3 - 2 = 3 - (3 - 1) = x - y(x - y) = (x - y)(x - y) = 2^2 = 4$.
11. Find the error in the following. $\frac{xy+y}{x} = y + y = 2y$. Now let $x = 2$ and $y = 2$ to obtain $3 = 4$.
12. Show the rational numbers satisfy the field axioms. You may assume the associative, commutative, and distributive laws hold for the integers.

1.3 Set Notation

A set is just a collection of things called elements. Often these are also referred to as points in calculus. For example $\{1, 2, 3, 8\}$ would be a set consisting of the elements 1, 2, 3, and 8. To indicate that 3 is an element of $\{1, 2, 3, 8\}$, it is customary to write $3 \in \{1, 2, 3, 8\}$. $9 \notin \{1, 2, 3, 8\}$ means 9 is not an element of $\{1, 2, 3, 8\}$. Sometimes a rule specifies a set. For example you could specify a set as all integers larger than 2. This would be written as $S = \{x \in \mathbb{Z} : x > 2\}$. This notation says: the set of all integers x , such that $x > 2$.

If A and B are sets with the property that every element of A is an element of B , then A is a subset of B . For example, $\{1, 2, 3, 8\}$ is a subset of $\{1, 2, 3, 4, 5, 8\}$, in symbols, $\{1, 2, 3, 8\} \subseteq \{1, 2, 3, 4, 5, 8\}$. The same statement about the two sets may also be written as $\{1, 2, 3, 4, 5, 8\} \supseteq \{1, 2, 3, 8\}$.

The union of two sets is the set consisting of everything which is contained in at least one of the sets, A or B . As an example of the union of two sets, $\{1, 2, 3, 8\} \cup \{3, 4, 7, 8\} = \{1, 2, 3, 4, 7, 8\}$ because these numbers are those which are in at least one of the two sets. Note that 3 is in both of these sets. In general

$$A \cup B \equiv \{x : x \in A \text{ or } x \in B\}.$$

Be sure you understand that something which is in both A and B is in the union. It is not an exclusive or.

The intersection of two sets, A and B consists of everything which is in both of the sets. Thus $\{1, 2, 3, 8\} \cap \{3, 4, 7, 8\} = \{3, 8\}$ because 3 and 8 are those elements the two sets have in common. In general,

$$A \cap B \equiv \{x : x \in A \text{ and } x \in B\}.$$

The symbol A^C indicates the set of things not in A . It makes sense when $A \subseteq U$, a universal set and it more precisely written as $U \setminus A$.

When \mathcal{K} is a set whose elements are sets, $\cap \mathcal{K}$ means everything which is in each of the sets of \mathcal{K} . Also $\cup \mathcal{K}$ is defined similarly. It is everything which is in at least one set of \mathcal{K} . More precisely, $\cap \mathcal{K} \equiv \cap \{K : K \in \mathcal{K}\}$. The following proposition is on De'Morgan's laws.

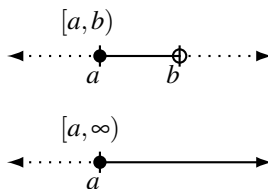
Proposition 1.3.1 *Let \mathcal{K} denote a set whose elements are subsets of some universal set U . Then*

$$(\cap \mathcal{K})^C = \cup \{K^C : K \in \mathcal{K}\}, (\cup \mathcal{K})^C = \cap \{K^C : K \in \mathcal{K}\}$$

Here $K^C \equiv U \setminus K$, everything outside of K .

Proof: This follows from the definition. To say $x \in (\cap \mathcal{K})^C$ is to say that x is not in the intersection of sets of \mathcal{K} which is to say that there is some set $K \in \mathcal{K}$ such that $x \notin K$ so $x \in K^C$ which is to say that $x \in \cup \{K^C : K \in \mathcal{K}\}$. The other claim is similar. ■

Intervals consist of sets of points on the real line which form a line segment or line. They might consist of all real numbers to the right of some point a including a and to the left of b not including b . This would be written as $[a, b)$. Maybe it is desired to specify all real numbers to the right of or equal to a . This would be written as $[a, \infty)$. Here are pictures of these two.



Other examples would be (a, b) which consists of all real numbers to the right of a not including a and also to the left of b . Note that this equals $(a, \infty) \cap (-\infty, b)$ where $(-\infty, b)$ denotes all real numbers left of b . In general, if the end point is included, you use $]$ or $[$ and if it is not included, you use $($ or $)$. This is the geometric description of intervals. In the next section, they are described in terms of order.

A special set which needs to be given a name is the empty set also called the null set, denoted by \emptyset . Thus \emptyset is defined as the set which has no elements in it. Mathematicians like to say the empty set is a subset of every set. The reason they say this is that if it were not so, there would have to exist a set, A , such that \emptyset has something in it which is not in A . However, \emptyset has nothing in it and so the least intellectual discomfort is achieved by saying $\emptyset \subseteq A$.

If A and B are two sets, $A \setminus B$ denotes the set of things which are in A but not in B . Thus $A \setminus B \equiv \{x \in A : x \notin B\}$. This is the same as $A \cap B^C$ where B^C indicates everything not in B . Set notation is used whenever convenient.

1.4 Order

Geometrically, order is defined as follows: $x < y$ means that y is right of x on the number line. This is also written as $y > x$. Also $y \geq x$ means y is to the right of x or maybe $y = x$. This is the way we usually think of order in calculus. However, there is a formal axiomatic description of order which follows. Most of these things are fairly obvious but I want to mention one especially. If $x < y$ and $a > 0$, then $ax < ay$ but in case $a < 0$, then $ax > ay$. You can see why this is from geometric reasoning. It would be good to convince yourself of this. Note that $-a = (-1)a$ and it simply reflects a across the 0 on the number line. The formal discussion follows.

The real numbers also have an order defined on them. This order may be defined by reference to the positive real numbers, those to the right of 0 on the number line, denoted by \mathbb{R}^+ , the positive numbers which is assumed to satisfy the following axioms.

Axiom 1.4.1 *The sum of two positive real numbers is positive.*

Axiom 1.4.2 *The product of two positive real numbers is positive.*

Axiom 1.4.3 *For a given real number x one and only one of the following alternatives holds. Either x is positive, $x = 0$, or $-x$ is positive.*

Definition 1.4.4 $x < y$ exactly when $y + (-x) \equiv y - x \in \mathbb{R}^+$. In the usual way, $x < y$ is the same as $y > x$ and $x \leq y$ means either $x < y$ or $x = y$. The symbol \geq is defined similarly.

Theorem 1.4.5 *The following hold for the order defined as above.*

1. If $x < y$ and $y < z$ then $x < z$ (Transitive law).
2. If $x < y$ then $x + z < y + z$ (addition to an inequality).
3. If $x \leq 0$ and $y \leq 0$, then $xy \geq 0$.
4. If $x > 0$ then $x^{-1} > 0$.
5. If $x < 0$ then $x^{-1} < 0$.
6. If $x < y$ then $xz < yz$ if $z > 0$, (multiplication of an inequality).
7. If $x < y$ and $z < 0$, then $xz > yz$ (multiplication of an inequality).
8. Each of the above holds with $>$ and $<$ replaced by \geq and \leq respectively except for 4 and 5 in which we must also stipulate that $x \neq 0$.
9. For any x and y , exactly one of the following must hold. Either $x = y$, $x < y$, or $x > y$ (trichotomy).

Proof: First consider 1, the transitive law. Suppose $x < y$ and $y < z$. Why is $x < z$? In other words, why is $z - x \in \mathbb{R}^+$? It is because $z - x = (z - y) + (y - x)$ and both $z - y, y - x \in \mathbb{R}^+$. Thus by 1.4.1 above, $z - x \in \mathbb{R}^+$ and so $z > x$.

Next consider 2, addition to an inequality. If $x < y$ why is $x + z < y + z$? it is because

$$\begin{aligned} (y + z) + -(x + z) &= (y + z) + (-1)(x + z) \\ &= y + (-1)x + z + (-1)z = y - x \in \mathbb{R}^+. \end{aligned}$$

Next consider 3. If $x \leq 0$ and $y \leq 0$, why is $xy \geq 0$? First note there is nothing to show if either x or y equal 0 so assume this is not the case. By 1.4.3 $-x > 0$ and $-y > 0$. Therefore, by 1.4.2 and what was proved about $-x = (-1)x$, $(-x)(-y) = (-1)^2 xy \in \mathbb{R}^+$. Is $(-1)^2 = 1$? If so the claim is proved. But $-(-1) = (-1)(-1) \equiv (-1)^2$ and $-(-1) = 1$ because $-1 + 1 = 0$. Therefore, $1 = (-1)^2$.

Next consider 4. If $x > 0$ why is $x^{-1} > 0$? By 1.4.3 either $x^{-1} = 0$ or $-x^{-1} \in \mathbb{R}^+$. It can't happen that $x^{-1} = 0$ because then you would have to have $1 = 0x$ and as was

shown earlier, $0x = 0$. Therefore, consider the possibility that $-x^{-1} \in \mathbb{R}^+$. This can't work either because then you would have $(-1)x^{-1}x = (-1)(1) = -1$ and it would follow from 1.4.2 that $-1 \in \mathbb{R}^+$. But this is impossible because if $x \in \mathbb{R}^+$, then $(-1)x = -x \in \mathbb{R}^+$ and contradicts 1.4.3 which states that either $-x$ or x is in \mathbb{R}^+ but not both.

Next consider 5. If $x < 0$, why is $x^{-1} < 0$? As before, $x^{-1} \neq 0$. If $x^{-1} > 0$, then as before, $-x(x^{-1}) = -1 \in \mathbb{R}^+$ which was just shown not to occur.

Next consider 6. If $x < y$ why is $xz < yz$ if $z > 0$? This follows because $yz - xz = z(y - x) \in \mathbb{R}^+$ since both z and $y - x \in \mathbb{R}^+$.

Next consider 7. If $x < y$ and $z < 0$, why is $xz > yz$? This follows because $zx - zy = z(x - y) \in \mathbb{R}^+$ by what was proved in 3.

The last two claims are obvious and left for you. This proves the theorem. ■

Note that trichotomy could be stated by saying $x \leq y$ or $y \leq x$. We say that two numbers x, y have the same sign if they are both in \mathbb{R}^+ or both $-x, -y$ are in \mathbb{R}^+ . A convenient way to tell is in the following proposition.

Proposition 1.4.6 *Two numbers x, y have the same sign if and only if $xy > 0$.*

Proof: \Rightarrow This follows from Axiom 1.4.2 if both are positive. If they are both negative, then $-x, -y$ are both positive and so $xy = (-x)(-y) > 0$.

\Leftarrow If $xy > 0$, then if $-x > 0$ and $y > 0$, then $-xy > 0$ so $xy < 0$. Hence $y < 0$. It is similar if $x > 0$. ■

Definition 1.4.7 $|x| \equiv \begin{cases} x & \text{if } x \geq 0, \\ -x & \text{if } x < 0. \end{cases}$

Note that $|x|$ can be thought of as the distance between x and 0.

Theorem 1.4.8 $|xy| = |x||y|$.

Proof: You can verify this by checking all available cases. Do so. You need consider both x, y nonnegative, both negative, and one negative and the other positive. ■

Theorem 1.4.9 *The following inequalities hold.*

$$|x + y| \leq |x| + |y|, \quad ||x| - |y|| \leq |x - y|.$$

Either of these inequalities may be called the triangle inequality.

Proof: First note that if $a, b \in \mathbb{R}^+ \cup \{0\}$ then $a \leq b$ if and only if $a^2 \leq b^2$. Here is why. Suppose $a \leq b$. Then by the properties of order proved above, $a^2 \leq ab \leq b^2$ because $b^2 - ab = b(b - a) \in \mathbb{R}^+ \cup \{0\}$. Next suppose $a^2 \leq b^2$. If both $a, b = 0$ there is nothing to show. Assume then they are not both 0. Then

$$b^2 - a^2 = (b + a)(b - a) \in \mathbb{R}^+ \cup \{0\}.$$

By the above theorem on order, $(a + b)^{-1} \in \mathbb{R}^+$ and so using the associative law,

$$(a + b)^{-1}((b + a)(b - a)) = (b - a) \in \mathbb{R}^+ \cup \{0\}$$

Thus $b \geq a$.

Now

$$\begin{aligned}|x+y|^2 &= (x+y)^2 = x^2 + 2xy + y^2 \\ &\leq |x|^2 + |y|^2 + 2|x||y| = (|x| + |y|)^2\end{aligned}$$

and so the first of the inequalities follows. Note I used $xy \leq |xy| = |x||y|$ which follows from the definition.

To verify the other form of the triangle inequality, $x = x - y + y$ so

$$|x| \leq |x - y| + |y|$$

and so $|x| - |y| \leq |x - y| = |y - x|$. Now repeat the argument replacing the roles of x and y to conclude $|y| - |x| \leq |y - x|$. Therefore,

$$||y| - |x|| \leq |y - x|.$$

This proves the triangle inequality. ■

Example 1.4.10 Solve the inequality $2x + 4 \leq x - 8$

Subtract $2x$ from both sides to yield $4 \leq -x - 8$. Next add 8 to both sides to get $12 \leq -x$. Then multiply both sides by (-1) to obtain $x \leq -12$. Alternatively, subtract x from both sides to get $x + 4 \leq -8$. Then subtract 4 from both sides to obtain $x \leq -12$.

Example 1.4.11 Solve the inequality $(x + 1)(2x - 3) \geq 0$.

If this is to hold, either both of the factors, $x + 1$ and $2x - 3$ are nonnegative or they are both non-positive. The first case yields $x + 1 \geq 0$ and $2x - 3 \geq 0$ so $x \geq -1$ and $x \geq \frac{3}{2}$ yielding $x \geq \frac{3}{2}$. The second case yields $x + 1 \leq 0$ and $2x - 3 \leq 0$ which implies $x \leq -1$ and $x \leq \frac{3}{2}$. Therefore, the solution to this inequality is $x \leq -1$ or $x \geq \frac{3}{2}$.

Example 1.4.12 Solve the inequality $(x)(x + 2) \geq -4$

Here the problem is to find x such that $x^2 + 2x + 4 \geq 0$. However, $x^2 + 2x + 4 = (x + 1)^2 + 3 \geq 0$ for all x . Therefore, the solution to this problem is all $x \in \mathbb{R}$.

Example 1.4.13 Solve the inequality $2x + 4 \leq x - 8$

This is written as $(-\infty, -12]$.

Example 1.4.14 Solve the inequality $(x + 1)(2x - 3) \geq 0$.

This was worked earlier and $x \leq -1$ or $x \geq \frac{3}{2}$ was the answer. In terms of set notation this is denoted by $(-\infty, -1] \cup [\frac{3}{2}, \infty)$.

Example 1.4.15 Solve the equation $|x - 1| = 2$

This will be true when $x - 1 = 2$ or when $x - 1 = -2$. Therefore, there are two solutions to this problem, $x = 3$ or $x = -1$.

Example 1.4.16 Solve the inequality $|2x - 1| < 2$

From the number line, it is necessary to have $2x - 1$ between -2 and 2 because the inequality says that the distance from $2x - 1$ to 0 is less than 2 . Therefore, $-2 < 2x - 1 < 2$ and so $-1/2 < x < 3/2$. In other words, $-1/2 < x$ and $x < 3/2$.

Example 1.4.17 Solve the inequality $|2x - 1| > 2$.

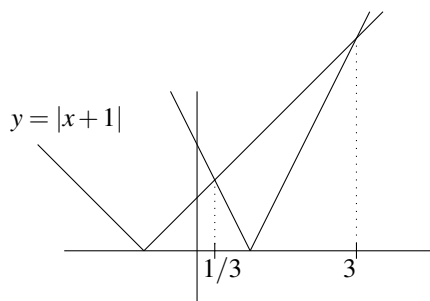
This happens if $2x - 1 > 2$ or if $2x - 1 < -2$. Thus the solution is $x > 3/2$ or $x < -1/2$. Written in terms of intervals this is $(\frac{3}{2}, \infty) \cup (-\infty, -\frac{1}{2})$.

Example 1.4.18 Solve $|x + 1| = |2x - 2|$

There are two ways this can happen. It could be the case that $x + 1 = 2x - 2$ in which case $x = 3$ or alternatively, $x + 1 = 2 - 2x$ in which case $x = 1/3$.

Example 1.4.19 Solve $|x + 1| \leq |2x - 2|$

In order to keep track of what is happening, it is a very good idea to graph the two relations, $y = |x + 1|$ and $y = |2x - 2|$ on the same set of coordinate axes. This is not a hard job. $|x + 1| = x + 1$ when $x > -1$ and $|x + 1| = -1 - x$ when $x \leq -1$. Therefore, it is not hard to draw its graph. Similar considerations apply to the other relation. The result is



Equality holds exactly when $x = 3$ or $x = \frac{1}{3}$ as in the preceding example. Consider x between $\frac{1}{3}$ and 3 . You can see these values of x do not solve the inequality. For example $x = 1$ does not work. Therefore, $(\frac{1}{3}, 3)$ must be excluded. The values of x larger than 3 do not produce equality so either $|x + 1| < |2x - 2|$ for these points or $|2x - 2| < |x + 1|$ for these points. Checking examples, you see the first of the two cases is the one which holds. Therefore, $[3, \infty)$ is included. Similar reasoning obtains $(-\infty, \frac{1}{3}]$. It follows the solution set to this inequality is $(-\infty, \frac{1}{3}] \cup [3, \infty)$.

Example 1.4.20 Suppose $\varepsilon > 0$ is a given positive number. Obtain a number, $\delta > 0$, such that if $|x - 1| < \delta$, then $|x^2 - 1| < \varepsilon$.

First of all, note $|x^2 - 1| = |x - 1||x + 1| \leq (|x| + 1)|x - 1|$. Now if $|x - 1| < 1$, it follows $|x| < 2$ and so for $|x - 1| < 1$,

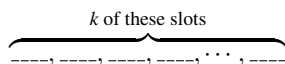
$$|x^2 - 1| < 3|x - 1|.$$

Now let $\delta = \min(1, \frac{\varepsilon}{3})$. This notation means to take the minimum of the two numbers, 1 and $\frac{\varepsilon}{3}$. Then if $|x - 1| < \delta$,

$$|x^2 - 1| < 3|x - 1| < 3\frac{\varepsilon}{3} = \varepsilon.$$

1. Solve $(3x+2)(x-3) \leq 0$.
2. Solve $(3x+2)(x-3) > 0$.
3. Solve $\frac{x+2}{3x-2} < 0$.
4. Solve $\frac{x+1}{x+3} < 1$.
5. Solve $(x-1)(2x+1) \leq 2$.
6. Solve $(x-1)(2x+1) > 2$.
7. Solve $x^2 - 2x \leq 0$.
8. Solve $(x+2)(x-2)^2 \leq 0$.
9. Solve $\frac{3x-4}{x^2+2x+2} \geq 0$.
10. Solve $\frac{3x+9}{x^2+2x+1} \geq 1$.
11. Solve $\frac{x^2+2x+1}{3x+7} < 1$.
12. Solve $|x+1| = |2x-3|$.
13. Solve $|3x+1| < 8$. Give your answer in terms of intervals on the real line.
14. Sketch on the number line the solution to the inequality $|x-3| > 2$.
15. Sketch on the number line the solution to the inequality $|x-3| < 2$.
16. Show $|x| = \sqrt{x^2}$.
17. Solve $|x+2| < |3x-3|$.
18. Tell when equality holds in the triangle inequality.
19. Solve $|x+2| \leq 8 + |2x-4|$.
20. Solve $(x+1)(2x-2)x \geq 0$.
21. Solve $\frac{x+3}{2x+1} > 1$.
22. Solve $\frac{x+2}{3x+1} > 2$.
23. Describe the set of numbers, a such that there is no solution to $|x+1| = 4 - |x+a|$.
24. Suppose $0 < a < b$. Show $a^{-1} > b^{-1}$.
25. Show that if $|x-6| < 1$, then $|x| < 7$.
26. Suppose $|x-8| < 2$. How large can $|x-5|$ be?
27. Obtain a number, $\delta > 0$, such that if $|x-1| < \delta$, then $|x^2-1| < 1/10$.
28. Obtain a number, $\delta > 0$, such that if $|x-4| < \delta$, then $|\sqrt{x}-2| < 1/10$.
29. Suppose $\varepsilon > 0$ is a given positive number. Obtain a number, $\delta > 0$, such that if $|x-1| < \delta$, then $|\sqrt{x}-1| < \varepsilon$. **Hint:** This δ will depend in some way on ε . You need to tell how.

Consider the following problem: You have the integers $S_n = \{1, 2, \dots, n\}$ and k is an integer no larger than n . How many ways are there to fill k slots with these integers starting from left to right if whenever an integer from S_n has been used, it cannot be re used in any succeeding slot?



This number is known as permutations of n things taken k at a time and is denoted by $P(n, k)$. It is easy to figure it out. There are n choices for the first slot. For each choice for the first slot, there remain $n - 1$ choices for the second slot. Thus there are $n(n - 1)$ ways to fill the first two slots. Now there remain $n - 2$ ways to fill the third. Thus there are

$n(n-1)(n-2)$ ways to fill the first three slots. Continuing this way, you see there are

$$P(n, k) = n(n-1)(n-2) \cdots (n-k+1)$$

ways to do this. Note there are k factors in the above product.

Now define for k a positive integer,

$$k! \equiv k(k-1)(k-2) \cdots 1, \quad 0! \equiv 1.$$

This is called k factorial. Thus $P(k, k) = k!$ and you should verify that

$$P(n, k) = \frac{n!}{(n-k)!}$$

Now consider the number of ways of selecting a set of k different numbers from S_n . For each set of k numbers there are $P(k, k) = k!$ ways of listing these numbers in order. Therefore, denoting by $\binom{n}{k}$ the number of ways of selecting a set of k numbers from S_n , it must be the case that

$$\binom{n}{k} k! = P(n, k) = \frac{n!}{(n-k)!}$$

Therefore, $\binom{n}{k} = \frac{n!}{k!(n-k)!}$. How many ways are there to select no numbers from S_n ?

Obviously one way. Note that the above formula gives the right answer in this case as well as in all other cases due to the definition which says $0! = 1$.

Now consider the problem of writing a formula for $(x+y)^n$ where n is a positive integer. Imagine writing it like this:

$$\overbrace{(x+y)(x+y) \cdots (x+y)}^{n \text{ times}}$$

Then you know the result will be sums of terms of the form $a_k x^k y^{n-k}$. What is a_k ? In other words, how many ways can you pick x from k of the factors above and y from the other $n-k$. There are n factors so the number of ways to do it is $\binom{n}{k}$. Therefore, a_k is the above formula and so this proves the following important theorem known as the binomial theorem.

Theorem 1.6.1 *The following formula holds for any n a positive integer.*

$$(x+y)^n = \sum_{k=0}^n \binom{n}{k} x^k y^{n-k}.$$

1.7 Well Ordering Principle, Math Induction

Definition 1.7.1 *A set is well ordered if every nonempty subset S , contains a smallest element z having the property that $z \leq x$ for all $x \in S$.*

Axiom 1.7.2 Any set of integers larger than a given number is well ordered.

In particular, the natural numbers defined as $\mathbb{N} \equiv \{1, 2, \dots\}$ is well ordered.

The above axiom implies the principle of mathematical induction.

Theorem 1.7.3 (Mathematical induction) A set $S \subseteq \mathbb{Z}$, having the property that $a \in S$ and $n + 1 \in S$ whenever $n \in S$ contains all integers $x \in \mathbb{Z}$ such that $x \geq a$.

Proof: Let $T \equiv ([a, \infty) \cap \mathbb{Z}) \setminus S$. Thus T consists of all integers larger than or equal to a which are not in S . The theorem will be proved if $T = \emptyset$. If $T \neq \emptyset$ then by the well ordering principle, there would have to exist a smallest element of T , denoted as b . It must be the case that $b > a$ since by definition, $a \notin T$. Then the integer, $b - 1 \geq a$ and $b - 1 \notin S$ because if $b - 1 \in S$, then $b - 1 + 1 = b \in S$ by the assumed property of S . Therefore, $b - 1 \in ([a, \infty) \cap \mathbb{Z}) \setminus S = T$ which contradicts the choice of b as the smallest element of T . ($b - 1$ is smaller.) Since a contradiction is obtained by assuming $T \neq \emptyset$, it must be the case that $T = \emptyset$ and this says that everything in $[a, \infty) \cap \mathbb{Z}$ is also in S . ■

Mathematical induction is a very useful device for proving theorems about the integers.

Example 1.7.4 Prove by induction that $\sum_{k=1}^n k^2 = \frac{n(n+1)(2n+1)}{6}$.

By inspection, if $n = 1$ then the formula is true. The sum yields 1 and so does the formula on the right. Suppose this formula is valid for some $n \geq 1$ where n is an integer. Then

$$\sum_{k=1}^{n+1} k^2 = \sum_{k=1}^n k^2 + (n+1)^2 = \frac{n(n+1)(2n+1)}{6} + (n+1)^2.$$

The step going from the first to the second line is based on the assumption that the formula is true for n . This is called the induction hypothesis. Now simplify the expression in the second line,

$$\frac{n(n+1)(2n+1)}{6} + (n+1)^2.$$

This equals $(n+1) \left(\frac{n(2n+1)}{6} + (n+1) \right)$ and

$$\frac{n(2n+1)}{6} + (n+1) = \frac{6(n+1) + 2n^2 + n}{6} = \frac{(n+2)(2n+3)}{6}$$

Therefore, $\sum_{k=1}^{n+1} k^2 = \frac{(n+1)(n+2)(2n+3)}{6} = \frac{(n+1)((n+1)+1)(2(n+1)+1)}{6}$, showing the formula holds for $n+1$ whenever it holds for n . This proves the formula by mathematical induction.

Example 1.7.5 Show that for all $n \in \mathbb{N}$, $\frac{1}{2} \cdot \frac{3}{4} \cdots \frac{2n-1}{2n} < \frac{1}{\sqrt{2n+1}}$.

If $n = 1$ this reduces to the statement that $\frac{1}{2} < \frac{1}{\sqrt{3}}$ which is obviously true. Suppose then that the inequality holds for n . Then

$$\frac{1}{2} \cdot \frac{3}{4} \cdots \frac{2n-1}{2n} \cdot \frac{2n+1}{2n+2} < \frac{1}{\sqrt{2n+1}} \cdot \frac{2n+1}{2n+2} = \frac{\sqrt{2n+1}}{2n+2}$$

The theorem will be proved if this last expression is less than $\frac{1}{\sqrt{2n+3}}$. This happens if and only if $\left(\frac{1}{\sqrt{2n+3}} \right)^2 = \frac{1}{2n+3} > \frac{2n+1}{(2n+2)^2}$ which occurs if and only if $(2n+2)^2 > (2n+3)(2n+1)$

and this is clearly true which may be seen from expanding both sides. This proves the inequality.

Lets review the process just used. If S is the set of integers at least as large as 1 for which the formula holds, the first step was to show $1 \in S$ and then that whenever $n \in S$, it follows $n + 1 \in S$. Therefore, by the principle of mathematical induction, S contains $[1, \infty) \cap \mathbb{Z}$, all positive integers. In doing an inductive proof of this sort, the set, S is normally not mentioned. One just verifies the steps above. First show the thing is true for some $a \in \mathbb{Z}$ and then verify that whenever it is true for m it follows it is also true for $m + 1$. When this has been done, the theorem has been proved for all $m \geq a$.

Definition 1.7.6 *The Archimedean property states that whenever $x \in \mathbb{R}$, and $a > 0$, there exists $n \in \mathbb{N}$ such that $na > x$.*

This is not hard to believe. Just look at the number line. Imagine the intervals

$$[0, a), [a, 2a), [2a, 3a), \dots$$

If $x < 0$, you could consider a and it would be larger than x . If $x \geq 0$, surely, it is reasonable to suppose that x would be on one of these intervals, say $[pa, (p + 1)a)$. This Archimedean property is quite important because it shows every fixed real number is smaller than some integer. It also can be used to verify a very important property of the rational numbers.

Axiom 1.7.7 \mathbb{R} has the Archimedean property.

Theorem 1.7.8 *Suppose $x < y$ and $y - x > 1$. Then there exists an integer, $l \in \mathbb{Z}$, such that $x < l < y$. If x is an integer, there is no integer y satisfying $x < y < x + 1$.*

Proof: Let x be the smallest positive integer. Not surprisingly, $x = 1$ but this can be proved. If $x < 1$ then $x^2 < x$ contradicting the assertion that x is the smallest natural number. Therefore, 1 is the smallest natural number. This shows there is no integer y , satisfying $x < y < x + 1$ since otherwise, you could subtract x and conclude $0 < y - x < 1$ for some integer $y - x$.

Now suppose $y - x > 1$ and let $S \equiv \{w \in \mathbb{N} : w \geq y\}$. The set S is nonempty by the Archimedean property. Let k be the smallest element of S . Therefore, $k - 1 < y$. Either

$k - 1 \leq x$ or $k - 1 > x$. If $k - 1 \leq x$, then $y - x \leq y - (k - 1) = \overbrace{y - k}^{\leq 0} + 1 \leq 1$ contrary to the assumption that $y - x > 1$. Therefore, $x < k - 1 < y$ and this proves the theorem with $l = k - 1$. ■

It is the next theorem which gives the density of the rational numbers. This means that for any real number, there exists a rational number arbitrarily close to it.

Theorem 1.7.9 *If $x < y$ then there exists a rational number r such that $x < r < y$.*

Proof: Let $n \in \mathbb{N}$ be large enough that $n(y - x) > 1$. Thus $(y - x)$ added to itself n times is larger than 1. Therefore, $n(y - x) = ny + n(-x) = ny - nx > 1$. It follows from Theorem 1.7.8 there exists $m \in \mathbb{Z}$ such that $nx < m < ny$ and so take $r = m/n$. ■

Definition 1.7.10 *A set, $S \subseteq \mathbb{R}$ is dense in \mathbb{R} if whenever $a < b$, $S \cap (a, b) \neq \emptyset$.*

Thus the above theorem says \mathbb{Q} is “dense” in \mathbb{R} .

You probably saw the process of division in elementary school. Even though you saw it at a young age it is very profound and quite difficult to understand. Suppose you want to do the following problem $\frac{79}{22}$. What did you do? You likely did a process of long division which gave the following result. $\frac{79}{22} = 3$ with remainder 13. This meant $79 = 3(22) + 13$. You were given two numbers, 79 and 22 and you wrote the first as some multiple of the second added to a third number which was smaller than the second number. Can this always be done? The answer is in the next theorem and depends here on the Archimedean property of the real numbers.

Theorem 1.7.11 *Suppose $0 < a$ and let $b \geq 0$. Then there exists a unique integer p and real number r such that $0 \leq r < a$ and $b = pa + r$.*

Proof: Let $S \equiv \{n \in \mathbb{N} : an > b\}$. By the Archimedean property this set is nonempty. Let $p + 1$ be the smallest element of S . Then $pa \leq b$ because $p + 1$ is the smallest in S . Therefore,

$$r \equiv b - pa \geq 0.$$

If $r \geq a$ then $b - pa \geq a$ and so $b \geq (p + 1)a$ contradicting $p + 1 \in S$. Therefore, $r < a$ as desired.

To verify uniqueness of p and r , suppose p_1 and $r_1, i = 1, 2$, both work and $r_2 > r_1$. Then a little algebra shows

$$p_1 - p_2 = \frac{r_2 - r_1}{a} \in (0, 1).$$

Thus $p_1 - p_2$ is an integer between 0 and 1, contradicting Theorem 1.7.8. The case that $r_1 > r_2$ cannot occur either by similar reasoning. Thus $r_1 = r_2$ and it follows that $p_1 = p_2$. ■

This theorem is called the Euclidean algorithm when a and b are integers. In this case, you would have r is an integer.

1.8 Exercises

1. By Theorem 1.7.9 it follows that for $a < b$, there exists a rational number between a and b . Show there exists an integer k such that $a < \frac{k}{2^m} < b$ for some k, m integers.
2. Show there is no smallest number in $(0, 1)$. Recall $(0, 1)$ means the real numbers which are strictly larger than 0 and smaller than 1.
3. Show there is no smallest number in $\mathbb{Q} \cap (0, 1)$.
4. Show that if $S \subseteq \mathbb{R}$ and S is well ordered with respect to the usual order on \mathbb{R} then S cannot be dense in \mathbb{R} .
5. Prove by induction that $\sum_{k=1}^n k^3 = \frac{1}{4}n^4 + \frac{1}{2}n^3 + \frac{1}{4}n^2$.
6. It is a fine thing to be able to prove a theorem by induction but it is even better to be able to come up with a theorem to prove in the first place. Derive a formula for $\sum_{k=1}^n k^4$ in the following way. Look for a formula in the form $An^5 + Bn^4 + Cn^3 +$

$Dn^2 + En + F$. Then try to find the constants A, B, C, D, E , and F such that things work out right. In doing this, show

$$(n+1)^4 = \left(A(n+1)^5 + B(n+1)^4 + C(n+1)^3 + D(n+1)^2 + E(n+1) + F \right) - \left(An^5 + Bn^4 + Cn^3 + Dn^2 + En + F \right),$$

so some progress can be made by matching the coefficients. When you get your answer, prove it is valid by induction.

7. Prove by induction that whenever $n \geq 2$, $\sum_{k=1}^n \frac{1}{\sqrt{k}} > \sqrt{n}$.
8. If $r \neq 0, 1$, show by induction that $\sum_{k=1}^n a(r^k) = a \frac{r^{n+1}}{r-1} - a \frac{r}{r-1}$.
9. Prove by induction that $\sum_{k=1}^n k = \frac{n(n+1)}{2}$.
10. Let a and d be real numbers. Find a formula for $\sum_{k=1}^n (a + kd)$ and then prove your result by induction.
11. Consider the geometric series, $\sum_{k=1}^n ar^{k-1}$. Prove by induction that if $r \neq 1$, then $\sum_{k=1}^n ar^{k-1} = \frac{a-ar^n}{1-r}$.
12. This problem is a continuation of Problem 11. You put money in the bank and it accrues interest at the rate of r per payment period. These terms need a little explanation. If the payment period is one month, and you started with \$100 then the amount at the end of one month would equal $100(1+r) = 100 + 100r$. In this the second term is the interest and the first is called the principal. Now you have $100(1+r)$ in the bank. How much will you have at the end of the second month? By analogy to what was just done it would equal $100(1+r) + 100(1+r)r = 100(1+r)^2$. In general, the amount you would have at the end of n months would be $100(1+r)^n$. (When a bank says they offer 6% compounded monthly, this means r , the rate per payment period equals .06/12.) In general, suppose you start with P and it sits in the bank for n payment periods. Then at the end of the n^{th} payment period, you would have $P(1+r)^n$ in the bank. In an ordinary annuity, you make payments, P at the end of each payment period, the first payment at the end of the first payment period. Thus there are n payments in all. Each accrue interest at the rate of r per payment period. Using Problem 11, find a formula for the amount you will have in the bank at the end of n payment periods? This is called the future value of an ordinary annuity. **Hint:** The first payment sits in the bank for $n-1$ payment periods and so this payment becomes $P(1+r)^{n-1}$. The second sits in the bank for $n-2$ payment periods so it grows to $P(1+r)^{n-2}$, etc.
13. Now suppose you want to buy a house by making n equal monthly payments. Typically, n is pretty large, 360 for a thirty year loan. Clearly a payment made 10 years from now can't be considered as valuable to the bank as one made today. This is because the one made today could be invested by the bank and having accrued interest for 10 years would be far larger. So what is a payment made at the end of k payment periods worth today assuming money is worth r per payment period? Shouldn't it be the amount, Q which when invested at a rate of r per payment period would yield P at the end of k payment periods? Thus from Problem 12 $Q(1+r)^k = P$ and so

$Q = P(1+r)^{-k}$. Thus this payment of P at the end of n payment periods, is worth $P(1+r)^{-k}$ to the bank right now. It follows the amount of the loan should equal the sum of these “discounted payments”. That is, letting A be the amount of the loan,

$$A = \sum_{k=1}^n P(1+r)^{-k}.$$

Using Problem 11, find a formula for the right side of the above formula. This is called the present value of an ordinary annuity.

14. Suppose the available interest rate is 7% per year and you want to take a loan for \$100,000 with the first monthly payment at the end of the first month. If you want to pay off the loan in 20 years, what should the monthly payments be? **Hint:** The rate per payment period is .07/12. See the formula you got in Problem 13 and solve for P .

15. Consider the first five rows of Pascal’s¹ triangle

$$\begin{array}{c} 1 \\ 1 \ 1 \\ 1 \ 2 \ 1 \\ 1 \ 3 \ 3 \ 1 \\ 1 \ 4 \ 6 \ 4 \ 1 \end{array}$$

What is the sixth row? Now consider that $(x+y)^1 = 1x + 1y$, $(x+y)^2 = x^2 + 2xy + y^2$, and $(x+y)^3 = x^3 + 3x^2y + 3xy^2 + y^3$. Give a conjecture about $(x+y)^5$.

16. Based on Problem 15 conjecture a formula for $(x+y)^n$ and prove your conjecture by induction. **Hint:** Letting the numbers of the n^{th} row of Pascal’s triangle be denoted by $\binom{n}{0}, \binom{n}{1}, \dots, \binom{n}{n}$ in reading from left to right, there is a relation between the numbers on the $(n+1)^{\text{st}}$ row and those on the n^{th} row, the relation being $\binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1}$. This is used in the inductive step.
17. Let $\binom{n}{k} \equiv \frac{n!}{(n-k)!k!}$ where $0! \equiv 1$ and $(n+1)! \equiv (n+1)n!$ for all $n \geq 0$. Prove that whenever $k \geq 1$ and $k \leq n$, then $\binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1}$. Are these numbers, $\binom{n}{k}$ the same as those obtained in Pascal’s triangle? Prove your assertion.
18. The binomial theorem states $(a+b)^n = \sum_{k=0}^n \binom{n}{k} a^{n-k} b^k$. Prove the binomial theorem by induction. **Hint:** You might try using the preceding problem.
19. Show that for $p \in (0, 1)$, $\sum_{k=0}^n \binom{n}{k} k p^k (1-p)^{n-k} = np$.
20. Using the binomial theorem prove that for all $n \in \mathbb{N}$, $\left(1 + \frac{1}{n}\right)^n \leq \left(1 + \frac{1}{n+1}\right)^{n+1}$. **Hint:** Show first that $\binom{n}{k} = \frac{n \cdot (n-1) \cdots (n-k+1)}{k!}$. By the binomial theorem,

$$\left(1 + \frac{1}{n}\right)^n = \sum_{k=0}^n \binom{n}{k} \left(\frac{1}{n}\right)^k = \sum_{k=0}^n \frac{\overbrace{n \cdot (n-1) \cdots (n-k+1)}^{k \text{ factors}}}{k! n^k}.$$

¹Blaise Pascal lived in the 1600’s and is responsible for the beginnings of the study of probability. He also did fundamental experiments on fluids.

Now consider the term $\frac{n \cdot (n-1) \cdots (n-k+1)}{k! n^k}$ and note that a similar term occurs in the binomial expansion for $\left(1 + \frac{1}{n+1}\right)^{n+1}$ except that n is replaced with $n+1$ wherever this occurs. Argue the term got bigger and then note that in the binomial expansion for $\left(1 + \frac{1}{n+1}\right)^{n+1}$, there are more terms.

21. Prove by induction that for all $k \geq 4$, $2^k \leq k!$
22. Use the Problems 21 and 20 to verify for all $n \in \mathbb{N}$, $\left(1 + \frac{1}{n}\right)^n \leq 3$.
23. Prove by induction that $1 + \sum_{i=1}^n i(i!) = (n+1)!$.
24. I can jump off the top of the Empire State Building without suffering any ill effects. Here is the proof by induction. If I jump from a height of one inch, I am unharmed. Furthermore, if I am unharmed from jumping from a height of n inches, then jumping from a height of $n+1$ inches will also not harm me. This is self evident and provides the induction step. Therefore, I can jump from a height of n inches for any n . What is the matter with this reasoning?
25. All horses are the same color. Here is the proof by induction. A single horse is the same color as himself. Now suppose the theorem that all horses are the same color is true for n horses and consider $n+1$ horses. Remove one of the horses and use the induction hypothesis to conclude the remaining n horses are all the same color. Put the horse which was removed back in and take out another horse. The remaining n horses are the same color by the induction hypothesis. Therefore, all $n+1$ horses are the same color as the $n-1$ horses which didn't get moved. This proves the theorem. Is there something wrong with this argument?
26. Let $\binom{n}{k_1, k_2, k_3}$ denote the number of ways of selecting a set of k_1 things, a set of k_2 things, and a set of k_3 things from a set of n things such that $\sum_{i=1}^3 k_i = n$. Find a formula for $\binom{n}{k_1, k_2, k_3}$. Now give a formula for a trinomial theorem, one which expands $(x+y+z)^n$. Could you continue this way and get a multinomial formula?

1.9 Completeness of \mathbb{R}

By Theorem 1.7.9, between any two real numbers, points on the number line, there exists a rational number. This suggests there are a lot of rational numbers, but it is not clear from this Theorem whether the entire real line consists of only rational numbers. Some people might wish this were the case because then each real number could be described, not just as a point on a line but also algebraically, as the quotient of integers. Before 500 B.C., a group of mathematicians, led by Pythagoras believed in this, but they discovered their beliefs were false. It happened roughly like this. They knew they could construct the square root of two as the diagonal of a right triangle in which the two sides have unit length; thus they could regard $\sqrt{2}$ as a number. Unfortunately, they were also able to show $\sqrt{2}$ could not be written as the quotient of two integers. This discovery that the rational numbers could not even account for the results of geometric constructions was very upsetting to the Pythagoreans, especially when it became clear there were an endless supply of such “irrational” numbers.

This shows that if it is desired to consider all points on the number line, it is necessary to abandon the attempt to describe arbitrary real numbers in a purely algebraic manner using only quotients of integers. Some might desire to throw out all the irrational numbers, and considering only the rational numbers, confine their attention to algebra, but this is not the approach to be followed here because it will effectively eliminate every major theorem of calculus. In this book real numbers will continue to be the points on the number line, a line which has no holes. This lack of holes is more precisely described in the following way.

Definition 1.9.1 *A non empty set, $S \subseteq \mathbb{R}$ is bounded above (below) if there exists $x \in \mathbb{R}$ such that $x \geq (\leq) s$ for all $s \in S$. If S is a nonempty set in \mathbb{R} which is bounded above, then a number, l which has the property that l is an upper bound and that every other upper bound is no smaller than l is called a least upper bound, l.u.b. (S) or often $\sup(S)$. If S is a nonempty set bounded below, define the greatest lower bound, g.l.b. (S) or $\inf(S)$ similarly. Thus g is the g.l.b. (S) means g is a lower bound for S and it is the largest of all lower bounds. If S is a nonempty subset of \mathbb{R} which is not bounded above, this information is expressed by saying $\sup(S) = +\infty$ and if S is not bounded below, $\inf(S) = -\infty$.*

Every existence theorem in calculus depends on some form of the completeness axiom. Bolzano was using this axiom in the early 1800's. It wasn't until late in that century that a construction of the real numbers from the rational numbers was completed by Dedekind. In this book, we will use this axiom whenever needed. Constructing the real numbers can be done later.

Axiom 1.9.2 (completeness) *Every nonempty set of real numbers which is bounded above has a least upper bound and every nonempty set of real numbers which is bounded below has a greatest lower bound.*

It is this axiom which distinguishes Calculus from Algebra. A fundamental result about \sup and \inf is the following.

Proposition 1.9.3 *Let S be a nonempty set and suppose $\sup(S)$ exists. Then for every $\delta > 0$,*

$$S \cap (\sup(S) - \delta, \sup(S)] \neq \emptyset.$$

If $\inf(S)$ exists, then for every $\delta > 0$,

$$S \cap [\inf(S), \inf(S) + \delta) \neq \emptyset.$$

Proof: Consider the first claim. If the indicated set equals \emptyset , then $\sup(S) - \delta$ is an upper bound for S which is smaller than $\sup(S)$, contrary to the definition of $\sup(S)$ as the least upper bound. In the second claim, if the indicated set equals \emptyset , then $\inf(S) + \delta$ would be a lower bound which is larger than $\inf(S)$ contrary to the definition of $\inf(S)$. ■

1.10 Existence of Roots

What is $\sqrt[5]{7}$ and does it even exist? You can ask for it on your calculator and the calculator will give you a number which multiplied by itself 5 times will yield a number which is close to 7 but it isn't exactly right. Why should there exist a number which works exactly?

Every one you find, appears to be some sort of approximation at best. If you can't produce one, why should you believe it is even there? Are you to accept it on faith like religion? Indeed, you must accept something without proof, but the appropriate thing to accept in the context of calculus is the completeness axiom of the real line on which every significant topic in calculus depends. In calculus, roots exist because of completeness of the real line as do integrals and all the major existence theorems in calculus, not because of algebraic techniques involving field extensions. Here is a lemma.

Lemma 1.10.1 *Suppose $n \in \mathbb{N}$ and $a > 0$. Then if $x^n - a \neq 0$, there exists $\delta > 0$ such that whenever $y \in (x - \delta, x + \delta)$, it follows $y^n - a \neq 0$ and has the same sign as $x^n - a$.*

Proof: From the binomial theorem, assuming always that $|y - x| < 1$,

$$\begin{aligned} y^n - a &= ((y - x) + x)^n - a = \sum_{k=0}^n \binom{n}{k} (y - x)^{n-k} x^k - a \\ &= \sum_{k=0}^{n-1} \binom{n}{k} (y - x)^{n-k} x^k + x^n - a = (y - x) \sum_{k=0}^{n-1} \binom{n}{k} (y - x)^{n-(k+1)} x^k + x^n - a \end{aligned}$$

Now from the triangle inequality and $|x - y| < 1$,

$$\begin{aligned} (x^n - a)(y^n - a) &= (x^n - a) \left((y - x) \sum_{k=0}^{n-1} \binom{n}{k} (y - x)^{n-(k+1)} x^k + (x^n - a) \right) \\ &\stackrel{=|x^n - a|^2}{\geq} (x^n - a)^2 - |y - x| |x^n - a| \sum_{k=0}^{n-1} \binom{n}{k} |x|^k \end{aligned}$$

Let $0 < \delta < \min \left(1, \frac{|x^n - a|}{2} \left(1 + \sum_{k=0}^{n-1} \binom{n}{k} |x|^k \right)^{-1} \right)$. Then if $|y - x| < \delta$, from the above,

$$(x^n - a)(y^n - a) \geq |x^n - a|^2 - \frac{|x^n - a|^2}{2} > 0$$

and so $x^n - a$ and $y^n - a$ have the same sign. ■

Theorem 1.10.2 *Let $a > 0$ and let $n > 1$. Then there exists a unique $x > 0$ such that $x^n = a$.*

Proof: Let S denote those numbers $y \geq 0$ such that $t^n - a < 0$ for all $t \in [0, y]$. Now note that from the binomial theorem,

$$(1 + a)^n - a = \sum_{k=0}^n \binom{n}{k} a^k 1^{n-k} - a \geq 1 + a - a = 1 > 0$$

Thus S is bounded above by $1 + a$ and $0 \in S$. Let $x \equiv \sup(S)$. Then by definition of sup, for every $\delta > 0$, there exists $t \in S$ with $|x - t| < \delta$.

If $x^n - a > 0$, then by the above lemma, for $t \in S$ sufficiently close to x ,

$$(t^n - a)(x^n - a) > 0$$

which is a contradiction because the first factor is negative and the second is positive. Hence $x^n - a \leq 0$. If $x^n - a < 0$, then from the above lemma, there is a $\delta > 0$ such that if $t \in (x - \delta, x + \delta)$, $x^n - a$ and $t^n - a$ have the same sign. This is also a contradiction because then $x \neq \sup(S)$. It follows $x^n = a$. ■

From now on, we will use this fact that n^{th} roots exist and are unique whenever it is convenient to do so.

1.11 Completing the Square

There is a very important process called completing the square. The idea is as follows. For a, b, c real numbers with $a \neq 0$, it is desired to write the expression $ax^2 + bx + c$ in the form $a(x - \gamma)^2 + \beta$. I will now show how to do this. It is very important because if you have done it, you can see that letting $x = \gamma$ yields the smallest possible value of the expression $ax^2 + bx + c$ for all x a real number provided $a > 0$ and it yields the largest possible value if $a < 0$. Here are the steps for doing it:

1. $ax^2 + bx + c = a\left(x^2 + \frac{b}{a}x + \frac{c}{a}\right)$
2. $a\left(x^2 + \frac{b}{a}x + \frac{c}{a}\right) = a\left(x^2 + \frac{b}{a}x + \frac{b^2}{4a^2} - \frac{b^2}{4a^2} + \frac{c}{a}\right)$
3. $a\left(x^2 + \frac{b}{a}x + \frac{b^2}{4a^2} - \frac{b^2}{4a^2} + \frac{c}{a}\right) = a\left(\left(x + \frac{b}{2a}\right)^2 - \left(\frac{b^2}{4a^2} - \frac{4ac}{4a^2}\right)\right)$
4. $a\left(\left(x + \frac{b}{2a}\right)^2 - \left(\frac{b^2}{4a^2} - \frac{4ac}{4a^2}\right)\right) = a\left(x + \frac{b}{2a}\right)^2 + \left(\frac{4ac - b^2}{4a}\right)$

The following fundamental theorem gives a formula for finding solutions to a quadratic equation, one of the form $ax^2 + bx + c = 0$.

Theorem 1.11.1 *If x is such that $ax^2 + bx + c = 0$ then*

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Proof: From the process of completing the square,

$$\left(x + \frac{b}{2a}\right)^2 = \left(\frac{b^2 - 4ac}{4a^2}\right)$$

and so on taking square roots, one obtains the two solutions described,

$$\frac{-b + \sqrt{b^2 - 4ac}}{2a}, \frac{-b - \sqrt{b^2 - 4ac}}{2a} \quad \blacksquare$$

Example 1.11.2 *Complete the square for $3x^2 + 4x - 5$.*

You can either go through the process or use the above formula. $a = 3, b = 4, c = -5$ and so, from the formula, this expression equals

$$3\left(x + \frac{4}{2(3)}\right)^2 + \left(\frac{4(3)(-5) - 4^2}{4(3)}\right) = 3\left(x + \frac{2}{3}\right)^2 + \left(-\frac{19}{3}\right)$$

Thus, in particular, the expression is minimized by letting $x = -\frac{2}{3}$ and its smallest value is $-\frac{19}{3}$.

Other situations are similar.

1.12 Dividing Polynomials

It will be very important to be able to work with polynomials, especially in subjects like linear algebra and with the technique of partial fractions. It is surprising how useful this junior high material will be. In this section, a polynomial is an expression. Later, the expression will be used to define a function. These two ways of looking at a polynomial are very different.

Definition 1.12.1 *A polynomial is an expression of the form*

$$p(\lambda) = a_n \lambda^n + a_{n-1} \lambda^{n-1} + \cdots + a_1 \lambda + a_0,$$

$a_n \neq 0$ where the a_i are numbers. Two polynomials are equal means that **the coefficients match for each power of λ** . The degree of a polynomial is the largest power of λ . Thus the degree of the above polynomial is n . Addition of polynomials is defined in the usual way as is multiplication of two polynomials. The leading term in the above polynomial is $a_n \lambda^n$. The coefficient of the leading term is called the leading coefficient. It is called a monic polynomial if $a_n = 1$. A **root** of a polynomial $p(\lambda)$ is μ such that $p(\mu) = 0$. This is also called a zero. When all coefficients are 0, we call it the zero polynomial.

The degree of the zero polynomial is not defined. Polynomials of degree 0 are the same as the numbers. The following is called the division algorithm. First is an important observation about multiplication of polynomials.

Lemma 1.12.2 *If $f(\lambda)g(\lambda) = 0$, then either $f(\lambda) = 0$ or $g(\lambda) = 0$. That is, there are no nonzero divisors of 0.*

Proof: Let $f(\lambda)$ have degree n and $g(\lambda)$ degree m . If $m+n=0$, it is easy to see that the conclusion holds. Suppose the conclusion holds for $m+n \leq M$ and suppose $m+n = M+1$. Then

$$\begin{aligned} 0 &= (a_0 + a_1 \lambda + \cdots + a_{n-1} \lambda^{n-1} + a_n \lambda^n) (b_0 + b_1 \lambda + \cdots + b_{m-1} \lambda^{m-1} + b_m \lambda^m) \\ &= (a(\lambda) + a_n \lambda^n) (b(\lambda) + b_m \lambda^m) \\ &= a(\lambda)b(\lambda) + b_m \lambda^m a(\lambda) + a_n \lambda^n b(\lambda) + a_n b_m \lambda^{n+m} \end{aligned}$$

Either $a_n = 0$ or $b_m = 0$ because their product is 0. Suppose $b_m = 0$. Then you need $(a(\lambda) + a_n \lambda^n)b(\lambda) = 0$. By induction, one of these polynomials in the product is 0. If $b(\lambda) \neq 0$, then this shows $a_n = 0$ and $a(\lambda) = 0$ so $f(\lambda) = 0$. If $b(\lambda) = 0$, then $g(\lambda) = 0$. The argument is the same if $a_n = 0$. ■

Lemma 1.12.3 *Let $f(\lambda)$ and $g(\lambda) \neq 0$ be polynomials. Then there exist polynomials, $q(\lambda)$ and $r(\lambda)$ such that*

$$f(\lambda) = q(\lambda)g(\lambda) + r(\lambda)$$

where the degree of $r(\lambda)$ is less than the degree of $g(\lambda)$ or $r(\lambda) = 0$. These polynomials $q(\lambda)$ and $r(\lambda)$ are unique.

Proof: Suppose that $f(\lambda) - q(\lambda)g(\lambda)$ is never equal to 0 for any $q(\lambda)$. If it is, then the conclusion follows. Now suppose

$$r(\lambda) = f(\lambda) - q(\lambda)g(\lambda) \tag{*}$$

where the degree of $r(\lambda)$ is as small as possible. Let it be m . Suppose $m \geq n$ where n is the degree of $g(\lambda)$. Say $r(\lambda) = b\lambda^m + a(\lambda)$ where $a(\lambda)$ is 0 or has degree less than m while $g(\lambda) = \hat{b}\lambda^n + \hat{a}(\lambda)$ where $\hat{a}(\lambda)$ is 0 or has degree less than n . Then

$$\begin{aligned} \left(f(\lambda) - \left(q(\lambda) + \frac{b}{\hat{b}}\lambda^{m-n} \right) g(\lambda) \right) &= r(\lambda) - \frac{b}{\hat{b}}\lambda^{m-n}g(\lambda) \\ &= b\lambda^m + a(\lambda) - \frac{b}{\hat{b}}\lambda^{m-n}(\hat{b}\lambda^n + \hat{a}(\lambda)) \end{aligned}$$

a polynomial having degree less than m . This is a contradiction because m was as small as possible and the left side is in the form of $*$. Hence $m < n$ after all.

As to uniqueness, if you have $r(\lambda), \hat{r}(\lambda), q(\lambda), \hat{q}(\lambda)$ which work, then you would have $(\hat{q}(\lambda) - q(\lambda))g(\lambda) = r(\lambda) - \hat{r}(\lambda)$. Now if the polynomial on the right is not zero, then neither is the one on the left. Hence this would involve two polynomials which are equal although their degrees are different. This is impossible. Hence $r(\lambda) = \hat{r}(\lambda)$ and so, the above lemma gives $\hat{q}(\lambda) = q(\lambda)$. ■

Definition 1.12.4 Let all coefficients of all polynomials come from a given field \mathbb{F} . For us, \mathbb{F} will be the real numbers \mathbb{R} . Let $p(\lambda) = a_n\lambda^n + \cdots + a_1\lambda + a_0$ be a polynomial. Recall it is called monic if $a_n = 1$. If you have polynomials

$$\{p_1(\lambda), \dots, p_m(\lambda)\},$$

the greatest common divisor $q(\lambda)$ is defined as the monic polynomial such that

1. $p_k(\lambda) = q(\lambda)l_k(\lambda)$ for some $l_k(\lambda)$ written as $q(\lambda)/p_k(\lambda)$ ($q(\lambda)$ divides $p_k(\lambda)$)
2. If $\hat{q}(\lambda)/p_k(\lambda)$ for each k , then $\hat{q}(\lambda)/q(\lambda)$.

A set of polynomials $\{p_1(\lambda), \dots, p_m(\lambda)\}$ is relatively prime if the greatest common divisor is 1.

Lemma 1.12.5 There is at most one greatest common divisor.

Proof: If you had two, $\hat{q}(\lambda)$ and $q(\lambda)$, then $\hat{q}(\lambda)/q(\lambda)$ and $q(\lambda)/\hat{q}(\lambda)$ so $q(\lambda) = \hat{q}(\lambda)\hat{l}(\lambda) = q(\lambda)l(\lambda)\hat{l}(\lambda)$ and now it follows, since both $\hat{q}(\lambda)$ and $q(\lambda)$ are monic that $\hat{l}(\lambda)$ and $l(\lambda)$ are both equal to 1. ■

The next proposition is remarkable. It describes the greatest common divisor in a very useful way.

Proposition 1.12.6 The greatest common divisor of $\{p_1(\lambda), \dots, p_m(\lambda)\}$ exists and is characterized as the monic polynomial of smallest degree equal to an expression of the form

$$\sum_{k=1}^m a_k(\lambda) p_k(\lambda), \text{ the } a_k(\lambda) \text{ being polynomials.} \quad (1.1)$$

Proof: First I need show that if $q(\lambda)$ is monic of the above form with smallest degree, then it is the greatest common divisor. If $q(\lambda)$ fails to divide $p_k(\lambda)$, then $p_k(\lambda) = q(\lambda)l(\lambda) + r(\lambda)$ where the degree of $r(\lambda)$ is smaller than the degree of $q(\lambda)$. Thus,

$$r(\lambda) = p_k(\lambda) - l(\lambda) \overbrace{\sum_{k=1}^m a_k(\lambda) p_k(\lambda)}^{q(\lambda)}$$

which violates the condition that $q(\lambda)$ has smallest degree because the right side is of the form in 1.1. Thus $q(\lambda) / p_k(\lambda)$ for each k . If $\hat{q}(\lambda)$ divides each $p_k(\lambda)$ then it must divide $q(\lambda)$ because $q(\lambda)$ is given by 1.1. Hence $q(\lambda)$ is the greatest common divisor.

Next, why does such greatest common divisor exist? Simply pick the monic polynomial which has smallest degree which is of the form $\sum_{k=1}^m a_k(\lambda) p_k(\lambda)$. Then from what was just shown, it is the greatest common divisor. ■

Proposition 1.12.7 *Let $p(\lambda)$ be a polynomial. Then there are polynomials $p_i(\lambda)$ such that*

$$p(\lambda) = a \prod_{i=1}^m p_i(\lambda)^{m_i} \quad (1.2)$$

where $m_i \in \mathbb{N}$ and $\{p_1(\lambda), \dots, p_m(\lambda)\}$ are monic and every subset of

$$\{p_1(\lambda), \dots, p_m(\lambda)\}$$

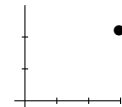
is relatively prime.

Proof: If there is no polynomial of degree larger than 0 dividing $p(\lambda)$, then we are done. Simply pick a such that $p(\lambda)$ is monic. Otherwise $p(\lambda) = ap_1(\lambda)p_2(\lambda)$ where $p_i(\lambda)$ is monic and each has degree at least 1. These could be the same polynomial. If some nonconstant polynomial divides either $p_i(\lambda)$, factor further. Continue doing this. Eventually the process must end with a factorization as described in 1.2 because the degrees of the factors are decreasing. Why is every subset of these $p_i(\lambda)$ relatively prime? If not, you could have factored the expression further. ■

1.13 The Complex Numbers

Recall that a real number is a point on the real number line. Just as a real number should be considered as a point on the line, a complex number is considered a point in the plane which can be identified in the usual way using the Cartesian coordinates of the point. Thus (a, b) identifies a point whose x coordinate is a and whose y coordinate is b . In dealing with complex numbers, such a point is written as $a + ib$. For example, in the following picture, I have graphed the point $3 + 2i$. You see it corresponds to the point in the plane whose coordinates are $(3, 2)$.

Multiplication and addition are defined in the most obvious way subject to the convention that $i^2 = -1$. Thus,



and

$$(a + ib) + (c + id) = (a + c) + i(b + d)$$

$$(a + ib)(c + id) = ac + iad + ibc + i^2bd = (ac - bd) + i(bc + ad).$$

Every non zero complex number $a + ib$, with $a^2 + b^2 \neq 0$, has a unique multiplicative inverse.

$$\frac{1}{a + ib} = \frac{a - ib}{a^2 + b^2} = \frac{a}{a^2 + b^2} - i \frac{b}{a^2 + b^2}.$$

You should prove the following theorem.

Theorem 1.13.1 *The complex numbers with multiplication and addition defined as above form a field satisfying all the field axioms. These are the following properties.*

1. $x + y = y + x$, (commutative law for addition)
2. $x + 0 = x$, (additive identity).
3. For each $x \in \mathbb{R}$, there exists $-x \in \mathbb{R}$ such that $x + (-x) = 0$, (existence of additive inverse).
4. $(x + y) + z = x + (y + z)$, (associative law for addition).
5. $xy = yx$, (commutative law for multiplication). You could write this as $x \times y = y \times x$.
6. $(xy)z = x(yz)$, (associative law for multiplication).
7. $1x = x$, (multiplicative identity).
8. For each $x \neq 0$, there exists x^{-1} such that $xx^{-1} = 1$. (existence of multiplicative inverse).
9. $x(y + z) = xy + xz$. (distributive law).

Something which satisfies these axioms is called a field. In this book, the field of most interest will be the field of real numbers. You have seen in earlier courses that the set of real numbers with the usual operations also satisfies the above axioms. The field of complex numbers is denoted as \mathbb{C} and the field of real numbers is denoted as \mathbb{R} . An important construction regarding complex numbers is the complex conjugate denoted by a horizontal line above the number. It is defined as follows.

$$\overline{a + ib} \equiv a - ib.$$

What it does is reflect a given complex number across the x axis. Algebraically, the following formula is easy to obtain.

$$(\overline{a + ib})(a + ib) = (a - ib)(a + ib) = a^2 + b^2 - i(ab - ab) = a^2 + b^2.$$

Observation 1.13.2 *The conjugate of a sum of complex numbers equals the sum of the complex conjugates and the conjugate of a product of complex numbers equals the product of the conjugates. To illustrate, consider the claim about the product.*

$$\begin{aligned} \overline{(a + ib)(c + id)} &= \overline{(ac - bd) + i(bc + ad)} = (ac - bd) - i(bc + ad) \\ \overline{(a + ib)} \overline{(c + id)} &= (a - ib)(c - id) = (ac - bd) - i(bc + ad) \end{aligned}$$

Showing the claim works for a sum is left for you. Of course this means the conclusion holds for any finite product or finite sum. Indeed, for z_k a complex number, the associative law of multiplication above gives

$$\overline{z_1 \cdots z_n} = \overline{(z_1 \cdots z_{n-1})(z_n)} = (\overline{z_1 \cdots z_{n-1}})(\overline{z_n})$$

Now by induction, the first product in the above can be split up into the product of the conjugates. Similar observations hold for sums.

Definition 1.13.3 *Define the absolute value of a complex number as follows.*

$$|a + ib| \equiv \sqrt{a^2 + b^2}.$$

Thus, denoting by z the complex number $z = a + ib$,

$$|z| = (z\overline{z})^{1/2}.$$

Also from the definition, if $z = x + iy$ and $w = u + iv$ are two complex numbers, then $|zw| = |z||w|$. You should verify this.

Notation 1.13.4 Recall the following notation. $\sum_{j=1}^n a_j \equiv a_1 + \cdots + a_n$. There is also a notation which is used to denote a product.

$$\prod_{j=1}^n a_j \equiv a_1 a_2 \cdots a_n$$

The triangle inequality holds for the absolute value for complex numbers just as it does for the ordinary absolute value.

Proposition 1.13.5 Let z, w be complex numbers. Then the triangle inequality holds.

$$|z + w| \leq |z| + |w|, \quad ||z| - |w|| \leq |z - w|.$$

Proof: Let $z = x + iy$ and $w = u + iv$. First note that

$$z\bar{w} = (x + iy)(u - iv) = xu + yv + i(yu - xv)$$

and so $|xu + yv| \leq |z\bar{w}| = |z||w|$.

$$\begin{aligned} |z + w|^2 &= (x + u + i(y + v))(x + u - i(y + v)) \\ &= (x + u)^2 + (y + v)^2 = x^2 + u^2 + 2xu + 2yv + y^2 + v^2 \\ &\leq |z|^2 + |w|^2 + 2|z||w| = (|z| + |w|)^2, \end{aligned}$$

so this shows the first version of the triangle inequality. To get the second,

$$z = z - w + w, \quad w = w - z + z$$

and so by the first form of the inequality

$$|z| \leq |z - w| + |w|, \quad |w| \leq |z - w| + |z|$$

and so both $|z| - |w|$ and $|w| - |z|$ are no larger than $|z - w|$ and this proves the second version because $||z| - |w||$ is one of $|z| - |w|$ or $|w| - |z|$. ■

With this definition, it is important to note the following. Be sure to verify this. It is not too hard but you need to do it.

Remark 1.13.6 : Let $z = a + ib$ and $w = c + id$. Then

$$|z - w| = \sqrt{(a - c)^2 + (b - d)^2}.$$

Thus the distance between the point in the plane determined by the ordered pair (a, b) and the ordered pair (c, d) equals $|z - w|$ where z and w are as just described.

For example, consider the distance between $(2, 5)$ and $(1, 8)$. From the distance formula this distance equals $\sqrt{(2 - 1)^2 + (5 - 8)^2} = \sqrt{10}$. On the other hand, letting $z = 2 + i5$ and $w = 1 + i8$, $z - w = 1 - i3$ and so $(z - w)(\overline{z - w}) = (1 - i3)(1 + i3) = 10$ so $|z - w| = \sqrt{10}$, the same thing obtained with the distance formula.

1.14 Polar Form of Complex Numbers

In the remaining sections of this chapter, I am assuming the reader knows basic trigonometry. If this is not the case, skip these sections and read them after the trig. functions have been developed systematically in the next chapter.

Complex numbers, are often written in the so called polar form which is described next. Suppose $z = x + iy$ is a complex number. Then

$$x + iy = \sqrt{x^2 + y^2} \left(\frac{x}{\sqrt{x^2 + y^2}} + i \frac{y}{\sqrt{x^2 + y^2}} \right).$$

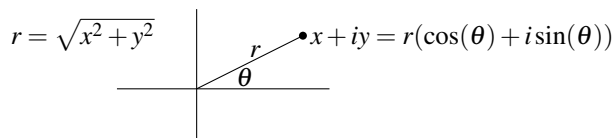
Now note that

$$\left(\frac{x}{\sqrt{x^2 + y^2}} \right)^2 + \left(\frac{y}{\sqrt{x^2 + y^2}} \right)^2 = 1$$

and so $\left(\frac{x}{\sqrt{x^2 + y^2}}, \frac{y}{\sqrt{x^2 + y^2}} \right)$ is a point on the unit circle. Therefore, there exists a unique angle $\theta \in [0, 2\pi)$ such that

$$\cos \theta = \frac{x}{\sqrt{x^2 + y^2}}, \quad \sin \theta = \frac{y}{\sqrt{x^2 + y^2}}.$$

The polar form of the complex number is then $r(\cos \theta + i \sin \theta)$ where θ is this angle just described and $r = \sqrt{x^2 + y^2} \equiv |z|$.



1.15 Roots of Complex Numbers

A fundamental identity is the formula of De Moivre which follows.

Theorem 1.15.1 *Let $r > 0$ be given. Then if n is a positive integer,*

$$[r(\cos t + i \sin t)]^n = r^n (\cos nt + i \sin nt).$$

Proof: It is clear the formula holds if $n = 1$. Suppose it is true for n .

$$[r(\cos t + i \sin t)]^{n+1} = [r(\cos t + i \sin t)]^n [r(\cos t + i \sin t)]$$

which by induction equals

$$\begin{aligned} &= r^{n+1} (\cos nt + i \sin nt) (\cos t + i \sin t) \\ &= r^{n+1} ((\cos nt \cos t - \sin nt \sin t) + i (\sin nt \cos t + \cos nt \sin t)) \\ &= r^{n+1} (\cos(n+1)t + i \sin(n+1)t) \end{aligned}$$

by the formulas for the cosine and sine of the sum of two angles. ■

Corollary 1.15.2 *Let z be a non zero complex number. Then there are always exactly k k^{th} roots of z in \mathbb{C} .*

Proof: Let $z = x + iy$ and let $z = |z|(\cos t + i \sin t)$ be the polar form of the complex number. By De Moivre's theorem, a complex number $r(\cos \alpha + i \sin \alpha)$, is a k^{th} root of z if and only if

$$r^k(\cos k\alpha + i \sin k\alpha) = |z|(\cos t + i \sin t).$$

This requires $r^k = |z|$ and so $r = |z|^{1/k}$ and also both $\cos(k\alpha) = \cos t$ and $\sin(k\alpha) = \sin t$. This can only happen if $k\alpha = t + 2l\pi$ for l an integer. Thus

$$\alpha = \frac{t + 2l\pi}{k}, l \in \mathbb{Z}$$

and so the k^{th} roots of z are of the form

$$|z|^{1/k} \left(\cos \left(\frac{t + 2l\pi}{k} \right) + i \sin \left(\frac{t + 2l\pi}{k} \right) \right), l \in \mathbb{Z}.$$

Since the cosine and sine are periodic of period 2π , there are exactly k distinct numbers which result from this formula. ■

Example 1.15.3 *Find the three cube roots of i .*

First note that $i = 1 \left(\cos \left(\frac{\pi}{2} \right) + i \sin \left(\frac{\pi}{2} \right) \right)$. Using the formula in the proof of the above corollary, the cube roots of i are

$$1 \left(\cos \left(\frac{(\pi/2) + 2l\pi}{3} \right) + i \sin \left(\frac{(\pi/2) + 2l\pi}{3} \right) \right)$$

where $l = 0, 1, 2$. Therefore, the roots are

$$\cos \left(\frac{\pi}{6} \right) + i \sin \left(\frac{\pi}{6} \right), \cos \left(\frac{5}{6}\pi \right) + i \sin \left(\frac{5}{6}\pi \right), \cos \left(\frac{3}{2}\pi \right) + i \sin \left(\frac{3}{2}\pi \right).$$

Thus the cube roots of i are $\frac{\sqrt{3}}{2} + i \left(\frac{1}{2} \right)$, $-\frac{\sqrt{3}}{2} + i \left(\frac{1}{2} \right)$, and $-i$.

The ability to find k^{th} roots can also be used to factor some polynomials.

Example 1.15.4 *Factor the polynomial $x^3 - 27$.*

First find the cube roots of 27. By the above procedure using De Moivre's theorem, these cube roots are $3, 3 \left(\frac{-1}{2} + i \frac{\sqrt{3}}{2} \right)$, and $3 \left(\frac{-1}{2} - i \frac{\sqrt{3}}{2} \right)$. Therefore, $x^3 - 27 =$

$$(x-3) \left(x-3 \left(\frac{-1}{2} + i \frac{\sqrt{3}}{2} \right) \right) \left(x-3 \left(\frac{-1}{2} - i \frac{\sqrt{3}}{2} \right) \right).$$

Note also $\left(x-3 \left(\frac{-1}{2} + i \frac{\sqrt{3}}{2} \right) \right) \left(x-3 \left(\frac{-1}{2} - i \frac{\sqrt{3}}{2} \right) \right) = x^2 + 3x + 9$ and so

$$x^3 - 27 = (x-3)(x^2 + 3x + 9)$$

where the quadratic polynomial $x^2 + 3x + 9$ cannot be factored without using complex numbers.

Note that even though the polynomial $x^3 - 27$ has all real coefficients, it has some complex zeros, $\frac{-1}{2} + i\frac{\sqrt{3}}{2}$ and $\frac{-1}{2} - i\frac{\sqrt{3}}{2}$. These zeros are complex conjugates of each other. It is **always** this way provided the coefficients of the polynomial are real. You should show this is the case. To see how to do this, see Problems 31 and 32 below.

Another fact for your information is the fundamental theorem of algebra. This theorem says that any polynomial of degree at least 1 having any complex coefficients always has a root in \mathbb{C} . This is sometimes referred to by saying \mathbb{C} is algebraically complete. Gauss is usually credited with giving a proof of this theorem in 1797 but many others worked on it and the first completely correct proof was due to Argand in 1806. For more on this theorem, you can google fundamental theorem of algebra and look at the interesting Wikipedia article on it. Proofs of this theorem usually involve the use of techniques from calculus even though it is really a result in algebra. A proof and plausibility explanation is given later.

1.16 Exercises

1. Let $S = [2, 5]$. Find $\sup S$. Now let $S = [2, 5)$. Find $\sup S$. Is $\sup S$ always a number in S ? Give conditions under which $\sup S \in S$ and then give conditions under which $\inf S \in S$.
2. Show that if $S \neq \emptyset$ and is bounded above (below) then $\sup S$ ($\inf S$) is unique. That is, there is only one least upper bound and only one greatest lower bound. If $S = \emptyset$ can you conclude that 7 is an upper bound? Can you conclude 7 is a lower bound? What about 13.5? What about any other number?
3. Let S be a set which is bounded above and let $-S$ denote the set $\{-x : x \in S\}$. How are $\inf(-S)$ and $\sup(S)$ related? **Hint:** Draw some pictures on a number line. What about $\sup(-S)$ and $\inf S$ where S is a set which is bounded below?
4. Which of the field axioms is being abused in the following argument that $0 = 2$? Let $x = y = 1$. Then $0 = x^2 - y^2 = (x - y)(x + y)$ and so $0 = (x - y)(x + y)$. Now divide both sides by $x - y$ to obtain $0 = x + y = 1 + 1 = 2$.
5. Give conditions under which equality holds in the triangle inequality.
6. Prove by induction that $n < 2^n$ for all natural numbers, $n \geq 1$.
7. Prove by the binomial theorem that the number of subsets of a given finite set containing n elements is 2^n .
8. Is it ever the case that $(a + b)^n = a^n + b^n$ for a and b positive real numbers?
9. Is it ever the case that $\sqrt{a^2 + b^2} = a + b$ for a and b positive real numbers?
10. Is it ever the case that $\frac{1}{x+y} = \frac{1}{x} + \frac{1}{y}$ for x and y positive real numbers?

11. Suppose $a > 0$ and that x is a real number which satisfies the quadratic equation, $ax^2 + bx + c = 0$. Go through the derivation given in the chapter for the quadratic formula. The expression $b^2 - 4ac$ is called the discriminant. When it is positive there are two different real roots. When it is zero, there is exactly one real root and when it equals a negative number there are no real roots. However, show that if x is given by the quadratic formula, it is in fact a solution to $ax^2 + bx + c = 0$ even though the square root will involve either of two complex numbers. **Hint:** You might observe that if a square root of a complex number z is w then the other square root is $-w$.
12. If α, β are roots of $x^2 + bx + c = 0$, then $(x - \alpha)(x - \beta) = 0$ so $x^2 - (\alpha + \beta)x + \alpha\beta = 0$ which means $-(\alpha + \beta) = b$ and so $-\frac{b}{2}$ is the average of the roots. Look for solutions in the form $-\frac{b}{2} + u$ and $-\frac{b}{2} - u$. Obtain the quadratic formula from this. ²
13. Suppose $f(x) = 3x^2 + 7x - 17$. Find the value of x at which $f(x)$ is smallest by completing the square. Also determine $f(\mathbb{R})$ and sketch the graph of f .
14. Suppose $f(x) = -5x^2 + 8x - 7$. Find $f(\mathbb{R})$. In particular, find the largest value of $f(x)$ and the value of x at which it occurs. Can you conjecture and prove a result about $y = ax^2 + bx + c$ in terms of the sign of a based on these last two problems?
15. Show that if it is assumed \mathbb{R} is complete, then the Archimedean property can be proved. **Hint:** Suppose completeness and let $a > 0$. If there exists $x \in \mathbb{R}$ such that $na \leq x$ for all $n \in \mathbb{N}$, then x/a is an upper bound for \mathbb{N} . Let l be the least upper bound and argue there exists $n \in \mathbb{N} \cap [l - 1/4, l]$. Now what about $n + 1$?
16. For those who know about the trigonometric functions, De Moivre's theorem says

$$[r(\cos t + i \sin t)]^n = r^n (\cos nt + i \sin nt)$$

for n a positive integer. Prove this formula by induction. Does this formula continue to hold for all integers n , even negative integers? Explain. **Hint:** I assume the reader knows the standard formulas for trig. functions like the sine and cosine of the sum of two variables. This is discussed later in the book. This problem is really about math induction.

17. Using De Moivre's theorem Theorem 1.15.1, derive a formula for $\sin(5x)$ and one for $\cos(5x)$. **Hint:** Use Problem 18 on Page 33 and if you like, you might use Pascal's triangle to construct the binomial coefficients.
18. De Moivre's theorem Theorem 1.15.1 is really a grand thing. I plan to use it now for rational exponents, not just integers.

$$1 = 1^{(1/4)} = (\cos 2\pi + i \sin 2\pi)^{1/4} = \cos(\pi/2) + i \sin(\pi/2) = i.$$

Therefore, squaring both sides it follows $1 = -1$. What does this tell you about De Moivre's theorem? Is there a profound difference between raising numbers to integer powers and raising numbers to non integer powers?

19. Review Problem 16 at this point. Now here is another question: If n is an integer, is it always true that $(\cos \theta - i \sin \theta)^n = \cos(n\theta) - i \sin(n\theta)$? Explain.

²The ancient Babylonians knew how to solve these quadratic equations sometime before 1700 B.C.

20. Suppose you have any polynomial in $\cos \theta$ and $\sin \theta$. By this I mean an expression of the form $\sum_{\alpha=0}^m \sum_{\beta=0}^n a_{\alpha\beta} \cos^\alpha \theta \sin^\beta \theta$ where $a_{\alpha\beta} \in \mathbb{R}$. Can this always be written in the form $\sum_{\gamma=-(n+m)}^{m+n} b_\gamma \cos \gamma\theta + \sum_{\tau=-(n+m)}^{n+m} c_\tau \sin \tau\theta$? Explain.
21. Let $z = 5 + i9$. Find z^{-1} .
22. Let $z = 2 + i7$ and let $w = 3 - i8$. Find $zw, z + w, z^2$, and w/z .
23. Give the complete solution to $x^4 + 16 = 0$.
24. Graph the complex cube roots of 8 in the complex plane. Do the same for the four fourth roots of 16.
25. If z is a complex number, show there exists ω a complex number with $|\omega| = 1$ and $\omega z = |z|$.
26. If z and w are two complex numbers and the polar form of z involves the angle θ while the polar form of w involves the angle ϕ , show that in the polar form for zw the angle involved is $\theta + \phi$. Also, show that in the polar form of a complex number z , $r = |z|$.
27. Factor $x^3 + 8$ as a product of linear factors.
28. Write $x^3 + 27$ in the form $(x + 3)(x^2 + ax + b)$ where $x^2 + ax + b$ cannot be factored any more using only real numbers.
29. Completely factor $x^4 + 16$ as a product of linear factors.
30. Factor $x^4 + 16$ as the product of two quadratic polynomials each of which cannot be factored further without using complex numbers.
31. If z, w are complex numbers prove $\overline{zw} = \overline{z}\overline{w}$ and then show by induction that $\overline{\prod_{j=1}^n z_j} = \prod_{j=1}^n \overline{z_j}$. Also verify that $\overline{\sum_{k=1}^m z_k} = \sum_{k=1}^m \overline{z_k}$. In words this says the conjugate of a product equals the product of the conjugates and the conjugate of a sum equals the sum of the conjugates.
32. Suppose $p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$ where all the a_k are real numbers. Suppose also that $p(z) = 0$ for some $z \in \mathbb{C}$. Show it follows that $p(\overline{z}) = 0$ also.
33. Show that $1 + i, 2 + i$ are the only two zeros to $p(x) = x^2 - (3 + 2i)x + (1 + 3i)$ so the zeros do not necessarily come in conjugate pairs if the coefficients are not real.
34. I claim that $1 = -1$. Here is why. $-1 = i^2 = \sqrt{-1}\sqrt{-1} = \sqrt{(-1)^2} = \sqrt{1} = 1$. This is clearly a remarkable result but is there something wrong with it? If so, what is wrong?
35. Suppose $p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$ is a polynomial and it has n zeros, z_1, z_2, \dots, z_n listed according to multiplicity. (z is a root of multiplicity m if the polynomial $f(x) = (x - z)^m$ divides $p(x)$ but $(x - z)f(x)$ does not.) Show that $p(x) = a_n (x - z_1)(x - z_2) \cdots (x - z_n)$.
36. Give the solutions to the following quadratic equations having real coefficients.

- (a) $x^2 - 2x + 2 = 0$ (d) $x^2 + 4x + 9 = 0$
 (b) $3x^2 + x + 3 = 0$
 (c) $x^2 - 6x + 13 = 0$ (e) $4x^2 + 4x + 5 = 0$

37. Give the solutions to the following quadratic equations having complex coefficients. Note how the solutions do not come in conjugate pairs as they do when the equation has real coefficients.

- (a) $x^2 + 2x + 1 + i = 0$ (d) $x^2 - 4ix - 5 = 0$
 (b) $4x^2 + 4ix - 5 = 0$
 (c) $4x^2 + (4 + 4i)x + 1 + 2i = 0$ (e) $3x^2 + (1 - i)x + 3i = 0$

38. Prove the fundamental theorem of algebra for quadratic polynomials having coefficients in \mathbb{C} . That is, show that an equation of the form $ax^2 + bx + c = 0$ where a, b, c are complex numbers, $a \neq 0$ has a complex solution. **Hint:** Consider the fact, noted earlier that the expressions given from the quadratic formula do in fact serve as solutions.

39. Suppose $r(\lambda) = \frac{a(\lambda)}{p(\lambda)^m}$ where $a(\lambda)$ is a polynomial and $p(\lambda)$ is an irreducible polynomial meaning that the only polynomials dividing $p(\lambda)$ are numbers and scalar multiples of $p(\lambda)$. That is, you can't factor it any further. Here we regard $r(\lambda)$ as a function. More on this later, but I assume people know about functions at this point. Show that $r(\lambda)$ is of the form

$$r(\lambda) = q(\lambda) + \sum_{k=1}^m \frac{b_k(\lambda)}{p(\lambda)^k}, \text{ where degree of } b_k(\lambda) < \text{degree of } p(\lambda)$$

40. \uparrow Suppose you have a rational function $\frac{a(\lambda)}{b(\lambda)}$.

- (a) Show it is of the form $p(\lambda) + \frac{n(\lambda)}{\prod_{i=1}^m p_i(\lambda)^{m_i}}$ where $\{p_1(\lambda), \dots, p_m(\lambda)\}$ are relatively prime and the degree of $n(\lambda)$ is less than the degree of $\prod_{i=1}^m p_i(\lambda)^{m_i}$.
 (b) Using Proposition 1.12.6 and the division algorithm for polynomials, show that the original rational function is of the form

$$\hat{p}(\lambda) + \sum_{i=1}^m \sum_{k=1}^{m_i} \frac{n_{ki}(\lambda)}{p_i(\lambda)^k}$$

where the degree of $n_{ki}(\lambda)$ is less than the degree of $p_i(\lambda)$ and $\hat{p}(\lambda)$ is some polynomial.

This is the partial fractions expansion of the rational function. Actually carrying out this computation may be impossible, but this shows the existence of such a partial fractions expansion.

41. In the above problem, use the fundamental theorem of algebra to show that for real polynomials, so all coefficients are in \mathbb{R} , the degree of each $p_i(\lambda)$ can always be taken no more than 2. For complex polynomials, the degree of each $p_i(\lambda)$ can be taken as 1. See Problem 32 above.

1.17 Videos

[Numbers](#) [Induction and binomial theorem](#) [polynomials and rational functions](#)

Chapter 2

Functions

2.1 General Considerations

The concept of a function is that of something which gives a unique output for a given input. This was likely first formulated in this way by Dirichlet. He wanted to consider piecewise continuous functions which were not given by a single formula. Often we think of functions in terms of formulas but the idea is more general and much older. In Luke 6:44, Jesus says essentially that you know a tree by its fruit. See also Matt. 7 about how to recognize false prophets. You look at what it does to determine what it is. As it is with people and trees, so it is with functions.

Definition 2.1.1 Consider two sets, D and R along with a rule which assigns a unique element of R to every element of D . This rule is called a function and it is denoted by a letter such as f . The symbol, $D(f) = D$ is called the domain of f . The set R , also written $R(f)$, is called the range of f . The set of all elements of R which are of the form $f(x)$ for some $x \in D$ is often denoted by $f(D)$. When $R = f(D)$, the function f is said to be onto. It is common notation to write $f : D(f) \rightarrow R$ to denote the situation just described in this definition where f is a function defined on D having values in R .

Example 2.1.2 Consider the list of numbers, $\{1, 2, 3, 4, 5, 6, 7\} \equiv D$. Define a function which assigns an element of D to $R \equiv \{2, 3, 4, 5, 6, 7, 8\}$ by $f(x) \equiv x + 1$ for each $x \in D$.

In this example there was a clearly defined procedure which determined the function. However, sometimes there is no discernible procedure which yields a particular function.

Example 2.1.3 Consider the ordered pairs, $(1, 2), (2, -2), (8, 3), (7, 6)$ and let the domain be given by $D \equiv \{1, 2, 8, 7\}$, the set of first entries in the given set of ordered pairs, $R \equiv \{2, -2, 3, 6\}$, the set of second entries, and let $f(1) = 2$, $f(2) = -2$, $f(8) = 3$, and $f(7) = 6$.

Sometimes functions are not given in terms of a formula. For example, consider the following function defined on the positive real numbers having the following definition.

Example 2.1.4 For $x \in \mathbb{R}$ define

$$f(x) = \begin{cases} \frac{1}{n} & \text{if } x = \frac{m}{n} \text{ in lowest terms for } m, n \in \mathbb{Z} \\ 0 & \text{if } x \text{ is not rational} \end{cases} \quad (2.1)$$

This is a very interesting function called the Dirichlet function. Note that it is not defined in a simple way from a formula.

Example 2.1.5 *My phone number has 10 digits. Let $f : \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\} \rightarrow \mathbb{N}$ be defined as follows. $f(k)$ is the k^{th} digit in my phone number. Thus $f(1) = 8$ because my area code starts with 8.*

This is not a very interesting function. I want to emphasize that functions are defined in terms of what they do rather than in terms of some formula although in calculus, we usually use functions which do come from a formula.

Example 2.1.6 *Let D consist of the set of people who have lived on the earth except for Adam and for $d \in D$, let $f(d) \equiv$ the biological father of d . Then f is a function.*

This function is not the sort of thing studied in calculus but it is a function just the same. When $D(f)$ is not specified, it is understood to consist of everything for which f makes sense. The following definition gives several ways to make new functions from old ones.

Definition 2.1.7 *Let f, g be functions with values in \mathbb{R} . Let a, b be points of \mathbb{R} . Then $af + bg$ is the name of a function whose domain is $D(f) \cap D(g)$ which is defined as*

$$(af + bg)(x) = af(x) + bg(x).$$

The function fg is the name of a function which is defined on $D(f) \cap D(g)$ given by

$$(fg)(x) = f(x)g(x).$$

Similarly for k an integer, f^k is the name of a function defined as

$$f^k(x) = (f(x))^k$$

The function f/g is the name of a function whose domain is

$$D(f) \cap \{x \in D(g) : g(x) \neq 0\}$$

defined as

$$(f/g)(x) = f(x)/g(x).$$

If $f : D(f) \rightarrow X$ and $g : D(g) \rightarrow Y$, then $g \circ f$ is the name of a function whose domain is

$$\{x \in D(f) : f(x) \in D(g)\}$$

which is defined as

$$g \circ f(x) \equiv g(f(x)).$$

This is called the composition of the two functions.

You should note that $f(x)$ is not a function. It is the value of the function at the point x . The name of the function is f . Nevertheless, people often write $f(x)$ to denote a function and it doesn't cause too many problems in beginning courses. When this is done, the variable x should be considered as a generic variable free to be anything in $D(f)$.

Example 2.1.8 Let $f(t) = t$ and $g(t) = 1 + t$. Then $fg : \mathbb{R} \rightarrow \mathbb{R}$ is given by

$$fg(t) = t(1 + t) = t + t^2.$$

Example 2.1.9 Let $f(t) = 2t + 1$ and $g(t) = \sqrt{1 + t}$. Then

$$g \circ f(t) = \sqrt{1 + (2t + 1)} = \sqrt{2t + 2}$$

for $t \geq -1$. If $t < -1$ the inside of the square root sign is negative so makes no sense. Therefore, $g \circ f : \{t \in \mathbb{R} : t \geq -1\} \rightarrow \mathbb{R}$.

Note that in this last example, it was necessary to fuss about the domain of $g \circ f$ because g is only defined for certain values of t .

Example 2.1.10 Let $f(t) = t^2$ for $t \in [0, 1]$ and let $g(t) = t^2$ for $t \in [0, 3]$. Then these are different functions because they have different domains.

The concept of a one to one function is very important. This is discussed in the following definition.

Definition 2.1.11 For any function $f : D(f) \subseteq X \rightarrow Y$, define the following set known as the inverse image of y .

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

There may be many elements in this set, but when there is always only one element in this set for all $y \in f(D(f))$, the function f is one to one sometimes written, $1-1$. Thus f is one to one, $1-1$, if whenever $f(x) = f(x_1)$, then $x = x_1$. If f is one to one, the inverse function f^{-1} is defined on $f(D(f))$ and $f^{-1}(y) = x$ where $f(x) = y$. Thus from the definition, $f^{-1}(f(x)) = x$ for all $x \in D(f)$ and $f(f^{-1}(y)) = y$ for all $y \in f(D(f))$. Defining id by $\text{id}(z) \equiv z$ this says $f \circ f^{-1} = \text{id}$ and $f^{-1} \circ f = \text{id}$. Note that this is sloppy notation because the two id are totally different functions.

Polynomials and rational functions are particularly easy functions to understand because they do come from a simple formula.

Definition 2.1.12 A function f is a polynomial if

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$$

where the a_i are real or complex numbers and n is a nonnegative integer. In this case the degree of the polynomial, $f(x)$ is n . Thus the degree of a polynomial is the largest exponent appearing on the variable.

f is a rational function if

$$f(x) = \frac{h(x)}{g(x)}$$

where h and g are polynomials.

For example, $f(x) = 3x^5 + 9x^2 + 7x + 5$ is a polynomial of degree 5 and

$$f(x) \equiv \frac{3x^5 + 9x^2 + 7x + 5}{x^4 + 3x + x + 1}$$

is a rational function.

Note that in the case of a rational function, the domain of the function might not be all of \mathbb{R} . For example, if $f(x) = \frac{x^2+8}{x+1}$, the domain of f would be all complex numbers not equal to -1 . Also, $f(x)$ is not a function. It is a function evaluated at x . The name of the function is f . Another thing which is often done is to denote the function in terms of an algorithm like

$$x \rightarrow \frac{x^2 - x + 1}{2x + 6}$$

This signifies the function f such that

$$f(x) = \frac{x^2 - x + 1}{2x + 6}.$$

Closely related to the definition of a function is the concept of the graph of a function.

Definition 2.1.13 *Given two sets, X and Y , the Cartesian product of the two sets, written as $X \times Y$, is assumed to be a set described as follows.*

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

\mathbb{R}^2 denotes the Cartesian product of \mathbb{R} with \mathbb{R} .

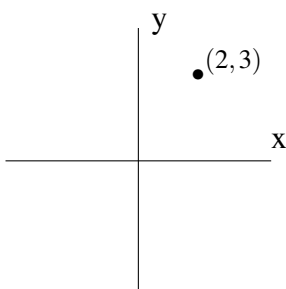
The notion of Cartesian product is just an abstraction of the concept of identifying a point in the plane with an ordered pair of numbers.

Definition 2.1.14 *Let $f : D(f) \rightarrow R(f)$ be a function. The graph of f consists of the set, $\{(x, y) : y = f(x) \text{ for } x \in D(f)\}$.*

Note that knowledge of the graph of a function is equivalent to knowledge of the function. To find $f(x)$, simply observe the ordered pair which has x as its first position on left and the value of y equals $f(x)$.

2.2 Graphs of Functions and Relations

Recall the notion of the Cartesian coordinate system you probably saw earlier. It involved an x axis, a y axis, two lines which intersect each other at right angles and one identifies a point by specifying a pair of numbers. For example, the number $(2, 3)$ involves going 2 units to the right on the x axis and then 3 units directly up on a line perpendicular to the x axis. For example, consider the following picture.



Because of the simple correspondence between points in the plane and the coordinates of a point in the plane, it is often the case that people are a little sloppy in referring to these things. Thus, it is common to see (x,y) referred to as a point in the plane. I will often indulge in this sloppiness. In terms of relations, if you graph the points as just described, you will have a way of visualizing the relation.

The reader has likely encountered the notion of graphing relations of the form $y = 2x + 3$ or $y = x^2 + 5$. The meaning of such an expression in terms of defining a relation is as follows. The relation determined by the equation $y = 2x + 3$ means the set of all ordered pairs (x,y) which are **related** by this formula. Thus the relation can be written as

$$\{(x,y) \in \mathbb{R}^2 : y = 2x + 3\}.$$

The relation determined by $y = x^2 + 5$ is $\{(x,y) \in \mathbb{R}^2 : y = x^2 + 5\}$. Note that these relations are also functions. For the first, you could let $f(x) = 2x + 3$ and this would tell you a rule which tells what the function does to x . However, some relations are not functions. For example, you could consider $x^2 + y^2 = 1$. Written more formally, the relation it defines is

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

Now if you give a value for x , there might be two values for y which are associated with the given value for x . In fact $y = \pm\sqrt{1-x^2}$. Thus this relation would not be a function.

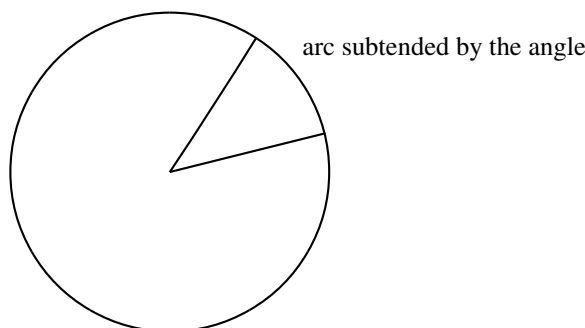
Recall how to graph a relation or more generally a relation. You first found lots of ordered pairs which satisfied the relation. For example $(0,3)$, $(1,5)$, and $(-1,1)$ all satisfy $y = 2x + 3$ which describes a straight line. Then you connected them with a curve.

2.3 Circular Functions

For a more thorough discussion of these functions along the lines given here, see my pre-calculus book published by Worldwide Center of Math. For a non geometric treatment, see my book Single variable advanced calculus or for a different way, Pure Mathematics by Hardy [19]. I much prefer methods which do not depend on plane geometry because with this approach, many of the most difficult and unpleasant considerations become obvious and then one can use the machinery of calculus to discuss geometric significance instead of relying so much on axioms from geometry which may or may not be well remembered. However, I am giving the traditional development of this subject here.

An angle consists of two lines emanating from a point as described in the following picture. How can angles be measured? This will be done by considering arcs on a circle. To see how this will be done, let θ denote an angle and place the vertex of this angle at the

center of the circle. Next, extend its two sides till they intersect the circle. Note the angle could be opening in any of infinitely many different directions. Thus this procedure could yield any of infinitely many different circular arcs. Each of these arcs is said to **subtend** the angle.



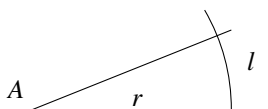
Take an angle and place its vertex (the point) at the center of a circle of radius r . Then, extending the sides of the angle if necessary till they intersect the circle, this determines an arc on the circle which subtends the angle. If r were changed to R , this really amounts to a change of units of length. Think, for example, of keeping the numbers the same but changing centimeters to meters in order to produce an enlarged version of the same picture. Thus the picture looks exactly the same, only larger. It is reasonable to suppose, based on this reasoning that the way to measure the angle is to take the length of the arc subtended in whatever units being used and divide this length by the radius measured in the same units, thus obtaining a number which is independent of the units of length used, just as the angle itself is independent of units of length. After all, it is the same angle regardless of how far its sides are extended. This is how to define the radian measure of an angle and the definition is well-defined. Thus, in particular, the ratio between the circumference (length) of a circle and its radius is a constant which is independent of the radius of the circle¹. Since the time of Euler in the 1700's, this constant has been denoted by 2π . In summary, if θ is the radian measure of an angle, the length of the arc subtended by the angle on a circle of radius r is $r\theta$.

So how do we obtain the length of the subtended arc? For now, imagine taking a string, placing one end of it on one end of the circular arc and then wrapping the string till you reach the other end of the arc. Stretching this string out and measuring it would then give you the length of the arc. Later a more precise way of finding lengths of curves is given.

Definition 2.3.1 *Let A be an angle. Draw a circle centered at A which intersects both sides of the angle. The radian measure of the angle is the length of this arc divided by the radius of the circle.*

¹In 2 Chronicles 4:2 the "molten sea" used for "washing" by the priests and found in Solomon's temple is described. It sat on 12 oxen, was round, 5 cubits high, 10 across and 30 around. This was very large if you believe what it says in Chronicles. A cubit is thought to have been about 1.5 feet. It is remarkable how much water was called for in their rituals. Their sacrifices also required a great deal of wood to burn up dead animals. This temple also exceeded the efficiency of a modern meat packing plant on some special occasions, according to the Bible.

Thus, from the above, the Bible gives the value of π as 3. This is not too far off and is much less pretentious than the Indiana pi bill of 1897 which attempted to legislate a method for squaring the circle. A better value is 3.1415926535 and presently this number is known to millions of decimal places. It was proved by Linderman in 1882 that π is transcendental which implies that it is impossible to construct a square having area π using only compass and unmarked straight edge (squaring the circle).

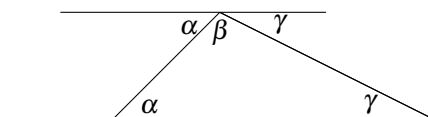


Thus the radian measure of A is l/r in the above. (Note that the radian measure of an angle does not depend on units of length.) There is also the **wrong** way of measuring angles. In this way, one degree consists of an angle which subtends an arc which goes $1/360$ of the way around the circle. The measure of the angle consists of the number of degrees which correspond to the given angle.

We avoid the wrong way of measuring angles in calculus. This is because all the theorems about the circular functions having to do with calculus topics involve the angle being given in radians.

In any triangle, the sum of the radian measures of the angles equals π . Let's review why this is so.

Consider the following picture.

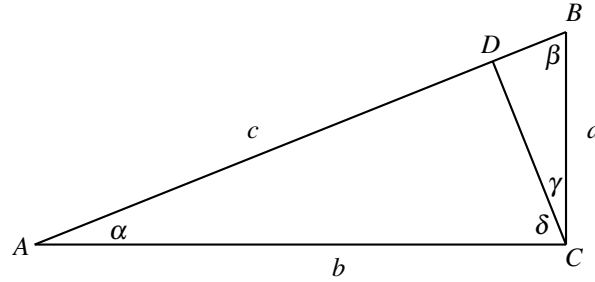


The line at the top is chosen to be parallel to the line on the base. Then from axioms of geometry about alternate interior angles, the diagram is correctly labeled. Now if you consider the angle formed by a point on a straight line, then it is obvious that the circle centered at this point has exactly half of it subtended by the line. Thus the radian measure of the angle is π . If we identify the radian measure of each of these angles with the label used for the angle, it follows that the sum of the measures of the angles of the triangle, $\alpha + \beta + \gamma$ equals π .

The following proof of the Pythagorean theorem is due to Euclid a few hundred years B.C. A right triangle is one in which one of the angles has radian measure $\pi/2$. It is called a right triangle. Thus if this angle is placed with its vertex at $(0,0)$ its sides subtend an arc of length $\pi/2$ on the unit circle, a circle with radius 1. The hypotenuse is by definition the side of the right triangle which is opposite the right angle. From the above observation, both of the other angles in a right triangle have radian measure less than $\pi/2$.

Theorem 2.3.2 (Pythagoras) *In a right triangle the square of the length of the hypotenuse equals the sum of the squares of the lengths of the other two sides.*

Proof: Consider the following picture in which the large triangle is a right triangle and D is the point where the line through C perpendicular to the line from A to B intersects the line from A to B . Then c is defined to be the length of the line from A to B , a is the length of the line from B to C , and b is the length of the line from A to C . Denote by \overline{DB} the length of the line from D to B .

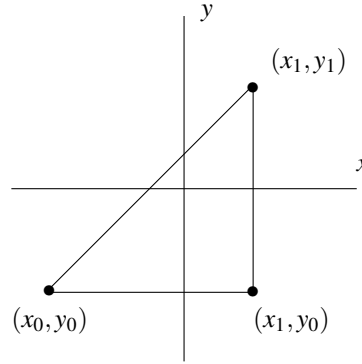


Then $\delta + \gamma = \pi/2$ and $\beta + \gamma = \pi/2$. Therefore, $\delta = \beta$. Also from this same theorem, $\alpha + \delta = \pi/2$ and so $\alpha = \gamma$. Therefore, the three triangles shown in the picture are all similar because they have the same angles at vertices. From the similar triangle axiom in geometry, the corresponding parts are proportional. Then

$$\frac{c}{a} = \frac{a}{\overline{DB}}, \text{ and } \frac{c}{b} = \frac{b}{c - \overline{DB}}.$$

Therefore, $c\overline{DB} = a^2$ and $c(c - \overline{DB}) = b^2$ so $c^2 = c\overline{DB} + b^2 = a^2 + b^2$. This proves the Pythagorean theorem. ² ■

Points in the plane may be identified by giving a pair of numbers. Suppose there are two points in the plane and it is desired to find the distance between them. There are actually many ways used to measure this distance but the best way, is determined by the Pythagorean theorem. Consider the following picture.



In this picture, the distance between the points denoted by (x_0, y_0) and (x_1, y_1) should be the square root of the sum of the squares of the lengths of the two sides. The length of the side on the bottom is $|x_0 - x_1|$ while the length of the side on the right is $|y_0 - y_1|$. Therefore, by the Pythagorean theorem the distance between the two indicated points is $\sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2}$. Note you could write

$$\sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2}$$

²This theorem is due to Pythagoras who lived about 572-497 B.C. This was during the Babylonian captivity of the Jews. Thus Pythagoras lived only a little more recently than Jeremiah. Nebuchadnezzar died a little after Pythagoras was born. Alexander the great would not come along for more than 100 years. There was, however, an even earlier Greek mathematician named Thales, 624-547 B.C. who also did fundamental work in geometry. Greek geometry was organized and published by Euclid about 300 B.C.

or even

$$\sqrt{(x_0 - x_1)^2 + (y_1 - y_0)^2}$$

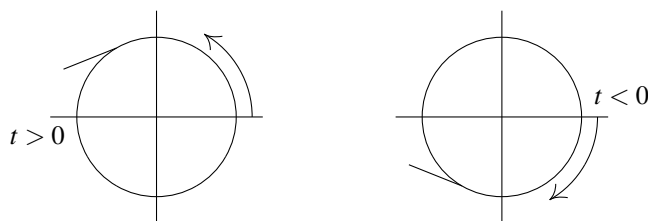
and it would make no difference in the resulting number. The distance between the two points is written as $|(x_0, y_0) - (x_1, y_1)|$ or sometimes when P_0 is the point determined by (x_0, y_0) and P_1 is the point determined by (x_1, y_1) , as $d(P_0, P_1)$ or $|P_0 P_1|$.

Thus, given an x and y axis at right angles to each other in the usual way, a relation which describes a point on the circle of radius 1 which has center at $(0, 0)$ is $x^2 + y^2 = 1$ or more precisely, $\{(x, y) : x^2 + y^2 = 1\}$.

This theorem implies there should exist some such number which deserves to be called $\sqrt{a^2 + b^2}$ as mentioned earlier in the discussion on completeness of \mathbb{R} .

Given a real number $t \in \mathbb{R}$, I will describe a point $p(t)$ on the unit circle.

Definition 2.3.3 Let $t \in \mathbb{R}$. If t is positive, take a string of length t , place one end at the point $(1, 0)$ and wrap the string **counter clockwise** around the circle which has radius 1 and center at $(0, 0)$ till you come to the end. This is the point $p(t)$. If t is negative, then take a string of length $|t|$ and with one end at $(1, 0)$, wrap the string in the **clockwise** direction around the unit circle till you come to the end. The point obtained is $p(t)$.



Definition 2.3.4 Let $t \in \mathbb{R}$. Then $p(t)$ will denote the point on the unit circle which was just described. Then $\sin(t)$ is the y coordinate of this point and $\cos(t)$ is the x coordinate of this point.

We say that $(\cos t, \sin t)$ parametrizes the unit circle with respect to arc length, but more on this much later. The thing to notice here is that a small change in t leads to a small change in $p(t)$ and consequently a small change in the x and y coordinates of $p(t)$ which are defined as $\cos t$ and $\sin t$ respectively. This is an informal way to say that these functions of t are continuous, but more about this will be discussed later.

Once you know about the sine and cosine, the other trigonometric functions are defined as follows.

$$\tan(x) \equiv \frac{\sin(x)}{\cos(x)}, \cot(x) \equiv \frac{\cos(x)}{\sin(x)}, \sec(x) \equiv \frac{1}{\cos(x)}, \csc(x) \equiv \frac{1}{\sin(x)}$$

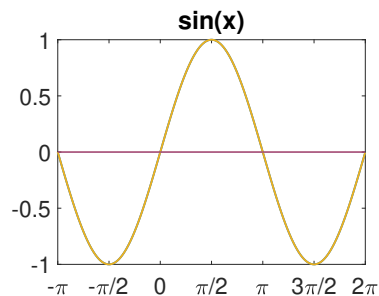
Of course those which are not the circular functions, sine and cosine, must have a restriction on their domains because one cannot divide by 0. Since $(\cos t, \sin t)$ is a point on the unit circle, it follows that

$$\cos^2 t + \sin^2 t = 1. \quad (2.2)$$

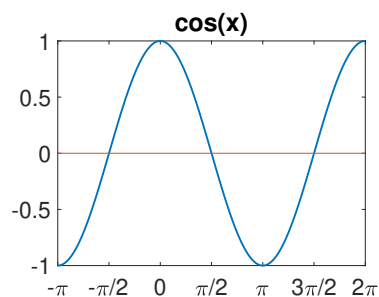
This is the most fundamental identity in trigonometry. Also, directly from the definition it follows that

$$\sin(t) = -\sin(-t), \cos(t) = \cos(-t) \quad (2.3)$$

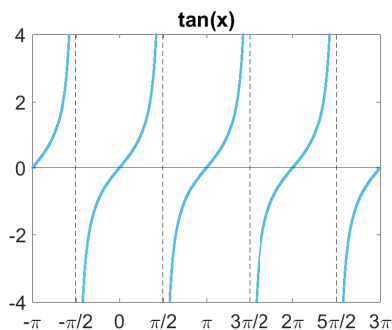
The above definitions are sufficient to determine approximately the values of the sine and cosine. Thus it is possible to produce a graph of these functions. Here is the graph of the function $y = \sin(x)$ on the interval $[-\pi, 2\pi]$. From the definition, the function is periodic of period 2π and so knowledge of the function on the interval $[0, 2\pi]$ is sufficient to describe it for all real values. Recall 2π is the length of the unit circle.



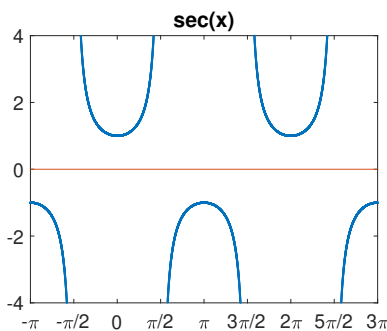
Now here is the graph of the function $y = \cos(x)$ on the interval $[-\pi, 2\pi]$. It is also periodic of period 2π .



As for the other functions, one can obtain graphs for them also. The function $x \rightarrow \tan(x)$ has the following graph on the interval $[-\pi, 3\pi]$. The vertical dashed lines are vertical asymptotes.



Finally, the graph of $x \rightarrow \sec(x)$ is of the form



Both of these functions have vertical asymptotes at odd multiples of $\pi/2$ although I have not shown them with the secant function.

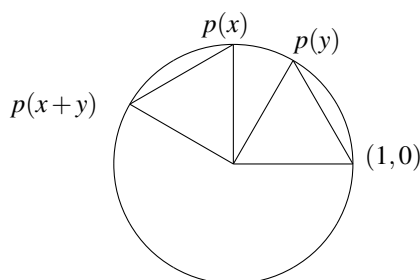
The formula for the cosine and sine of the sum of two angles is also important. Like most of this material, I assume the reader has seen it. However, I am aware that many people do not see these extremely important formulas, or if they do, they often see no explanation for them so I shall give a review of it here.

The following theorem is the fundamental identity from which all the major trig. identities involving sums and differences of angles are derived.

Theorem 2.3.5 *Let $x, y \in \mathbb{R}$. Then*

$$\cos(x+y)\cos(x) + \sin(x+y)\sin(x) = \cos(y). \quad (2.4)$$

Proof: Recall that for a real number z , there is a unique point $p(z)$ on the unit circle and the coordinates of this point are $\cos z$ and $\sin z$. Now it seems geometrically clear that the length of the arc between $p(x+y)$ and $p(x)$ has the same length as the arc between $p(y)$ and $p(0)$. As in the following picture.



Also from geometric reasoning, rigorously examined later, the distance between the points $p(x+y)$ and $p(x)$ must be the same as the distance from $p(y)$ to $p(0)$. In fact, the two triangles have the same angles and the same sides. Writing this in terms of the definition of the trig functions and the distance formula,

$$(\cos(x+y) - \cos x)^2 + (\sin(x+y) - \sin x)^2 = (\cos y - 1)^2 + \sin^2 y.$$

Expanding the above,

$$\begin{aligned} \cos^2(x+y) + \cos^2 x - 2\cos(x+y)\cos x + \sin^2(x+y) + \sin^2 x - 2\sin(x+y)\sin x \\ = \cos^2 y - 2\cos y + 1 + \sin^2 y \end{aligned}$$

Now using that $\cos^2 + \sin^2 = 1$,

$$2 - 2\cos(x+y)\cos(x) - 2\sin(x+y)\sin(x) = 2 - 2\cos(y).$$

Therefore,

$$\cos(x+y)\cos(x) + \sin(x+y)\sin(x) = \cos(y) \blacksquare$$

2.3.1 Reference Angles and Other Identities

Recall that the length of the unit circle is defined as 2π . This started with Euler who decided that π should be such that 2π is the length of the unit circle. Thus it becomes obvious what the sine and cosine are for certain special angles. For example, $\sin(\frac{\pi}{2}) = 1$, $\cos(\frac{\pi}{2}) = 0$. Letting $x = \pi/2$, 2.4 shows that

$$\sin(y + \pi/2) = \cos y. \quad (2.5)$$

Now let $u = x + y$ and $v = x$. Then 2.4 implies

$$\cos u \cos v + \sin u \sin v = \cos(u - v) \quad (2.6)$$

Also, from this and 2.3,

$$\begin{aligned} \cos(u + v) &= \cos(u - (-v)) = \cos u \cos(-v) + \sin u \sin(-v) \\ &= \cos u \cos v - \sin u \sin v \end{aligned} \quad (2.7)$$

Thus, letting $v = \pi/2$,

$$\cos\left(u + \frac{\pi}{2}\right) = -\sin u. \quad (2.8)$$

It follows

$$\begin{aligned} \sin(x + y) &= -\cos\left(x + \frac{\pi}{2} + y\right) \\ &= -\left[\cos\left(x + \frac{\pi}{2}\right)\cos y - \sin\left(x + \frac{\pi}{2}\right)\sin y\right] \\ &= \sin x \cos y + \sin y \cos x \end{aligned} \quad (2.9)$$

Then using 2.3, that $\sin(-y) = -\sin(y)$ and $\cos(-x) = \cos(x)$, this implies

$$\sin(x - y) = \sin x \cos y - \cos x \sin y. \quad (2.10)$$

In addition to this,

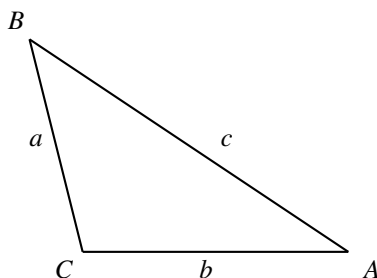
$$\cos 2x = \cos^2 x - \sin^2 x = 2\cos^2 x - 1 = 1 - 2\sin^2 x \quad (2.11)$$

Therefore, making use of the above identities,

$$\begin{aligned} \cos(3x) &= \cos(2x + x) = \cos 2x \cos x - \sin 2x \sin x \\ &= (2\cos^2 x - 1)\cos x - 2\cos x \sin^2 x \\ &= 4\cos^3 x - 3\cos x \end{aligned} \quad (2.12)$$

For a systematic way to find cosine or sine of a multiple of x , see De Moivre's theorem explained in Problem 16 on Page 46.

Another very important theorem from Trigonometry is the law of cosines. Consider the following picture of a triangle in which a, b and c are the lengths of the sides and A, B , and C denote the angles indicated.



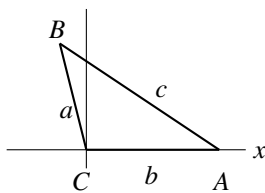
The law of cosines is the following.

Theorem 2.3.6 *Let ABC be a triangle as shown above. Then*

$$c^2 = a^2 + b^2 - 2ab \cos C$$

Also, $c \leq a + b$ so the length of a side of a triangle is no more than the sum of the lengths of the other two sides.

Proof: Situate the triangle so the vertex of the angle C , is on the point whose coordinates are $(0, 0)$ and so the side opposite the vertex B is on the positive x axis.



Then from the definition of the $\cos C$, the coordinates of the vertex B are

$$(a \cos C, a \sin C)$$

while it is clear that the coordinates of A are $(b, 0)$. Therefore, from the distance formula,

$$\begin{aligned} c^2 &= (a \cos C - b)^2 + a^2 \sin^2 C \\ &= a^2 \cos^2 C - 2ab \cos C + b^2 + a^2 \sin^2 C \\ &= a^2 + b^2 - 2ab \cos C \end{aligned}$$

For the last claim, $c^2 = a^2 + b^2 - 2ab \cos C \leq a^2 + b^2 + 2ab = (a + b)^2$. ■

As mentioned, you can find sine and cosine of certain special angles. However, all that was considered was $\pi/2$, but various other angles are easy to figure out also.

Example 2.3.7 Find $\cos(\pi/6)$.

Using 2.12 and the observation that $3\left(\frac{\pi}{6}\right) = \frac{\pi}{2}$,

$$0 = \cos\left(\frac{\pi}{2}\right) = 4\cos^3\left(\frac{\pi}{6}\right) - 3\cos\left(\frac{\pi}{6}\right)$$

and you can solve this for $\cos\left(\frac{\pi}{6}\right)$

$$4\cos^2\left(\frac{\pi}{6}\right) = 3, \quad \cos\left(\frac{\pi}{6}\right) = \frac{\sqrt{3}}{2}$$

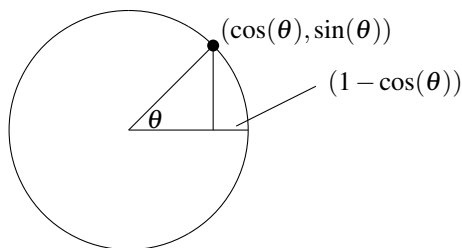
Example 2.3.8 Find $\cos(7\pi/6)$.

If you sketch where the point determined by $7\pi/6$ is, you see that this should be $-\cos(\pi/6) = -\sqrt{3}/2$. Other examples are done similarly.

I assume the reader has done these sorts of things so I will not belabor this much more. Good practice will be found in the exercises.

2.3.2 The $\sin(x)/x$ Inequality

There is an amazingly important inequality which to most of us is fairly obvious from the picture. Here is a picture which illustrates the conclusion of this corollary.



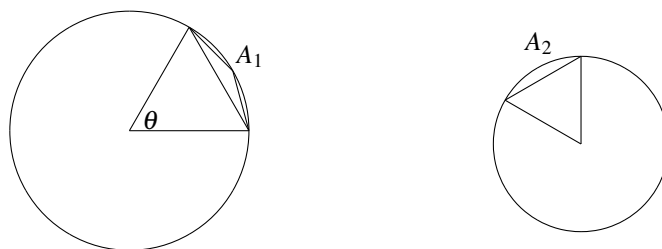
The corollary states that the length of the subtended arc shown in the picture is longer than the vertical side of the triangle and smaller than the sum of the vertical side with the segment having length $1 - \cos \theta$.

Corollary 2.3.9 Let $0 \leq \text{radian measure of } \theta < \pi/4$. Then letting A be the arc on the unit circle resulting from situating the angle with one side on the positive x axis and the other side pointing up from the positive x axis,

$$(1 - \cos \theta) + \sin \theta \geq l(A) \geq \sin \theta \quad (2.13)$$

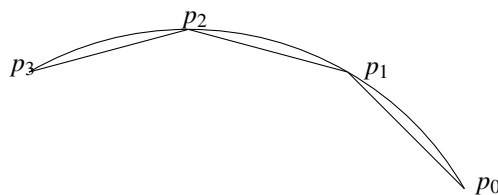
While this seems obvious and in fact you could easily convince yourself of its truth by graphing circles and using string, to do this right, one must give a more mathematically precise treatment of arc length on the circle. What exactly do we mean by “arc length”? First note that for $t \geq 0$, there is a unique nonnegative integer n such that $t = 2\pi n + l$ where $l \in [0, 2\pi)$. Similarly if $t < 0$ there is a unique nonnegative integer n such that $t = (-2\pi)n + l$ where again $l \in [0, 2\pi)$. Then to get to the point $p(t)$, one starts at $(0,0)$ and on the unit circle and moves in the counter clockwise direction a distance of l using the description of length of a circular arc about to be presented. Thus it suffices to consider the length of an arc on the unit circle or more generally an arc on a circle of radius r .

To give a precise description of what is meant by the length of an arc, consider the following picture.



In this picture, there are two circles, a big one having radius R and a little one having radius r . The angle θ is situated in two different ways subtending the arcs A_1 and A_2 as shown.

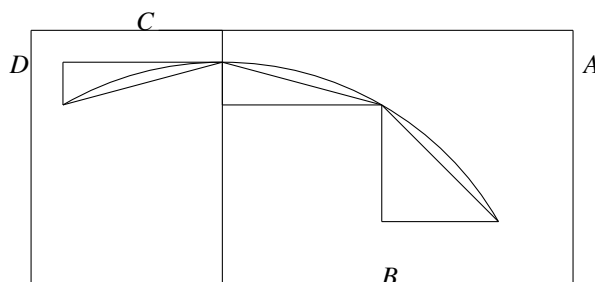
Letting A be an arc of a circle, like those shown in the above picture, a subset of $A, \{p_0, \dots, p_n\}$ is a partition of A if p_0 is one endpoint, p_n is the other end point, and the points are encountered in the indicated order as one moves in the counter clockwise direction along the arc. To illustrate, see the following picture.



Also, denote by $\mathcal{P}(A)$ the set of all such partitions. For $P = \{p_0, \dots, p_n\}$, denote by $|p_i - p_{i-1}|$ the distance between p_i and p_{i-1} . Then for $P \in \mathcal{P}(A)$, define

$$|P| \equiv \sum_{i=1}^n |p_i - p_{i-1}|$$

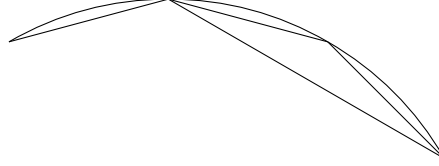
Thus $|P|$ consists of the sum of the lengths of the little lines joining successive points of P and appears to be an approximation to the length of the circular arc A . By Theorem 2.3.6 the length of any of the straight line segments joining successive points in a partition is smaller than the sum of the two sides of a right triangle having the given straight line segment as its hypotenuse. Now consider the following picture.



The sum of the lengths of the straight line segments in the part of the picture found in the right rectangle above is less than $A + B$ and the sum of the lengths of the straight line segments in the part of the picture found in the left rectangle above is less than $C + D$ and this would be so for any partition. Therefore, for any $P \in \mathcal{P}(A)$, $|P| \leq M$ where M is the perimeter of a rectangle containing the arc A . To be a little sloppy, simply pick M to be the perimeter of a rectangle containing the whole circle of which A is a part. The only purpose

for doing this is to obtain the existence of an upper bound. Therefore, $\{|P| : P \in \mathcal{P}(A)\}$ is a set of numbers which is bounded above by M and by completeness of \mathbb{R} it is possible to define the length of A , $l(A)$, by $l(A) \equiv \sup\{|P| : P \in \mathcal{P}(A)\}$.

A fundamental observation is that if $P, Q \in \mathcal{P}(A)$ and $P \subseteq Q$, then $|P| \leq |Q|$. To see this, add in one point at a time to P . This effect of adding in one point is illustrated in the following picture.



Remember Theorem 2.3.6 that the length of a side of a triangle is no more than the sum of the lengths of the other two sides.

Also, letting $\{p_0, \dots, p_n\}$ be a partition of A , specify angles, θ_i as follows. The angle θ_i is formed by the two lines, one from the center of the circle to p_i and the other line from the center of the circle to p_{i-1} . Furthermore, a specification of these angles yields the partition of A in the following way. Place the vertex of θ_1 on the center of the circle, letting one side lie on the line from the center of the circle to p_0 and the other side extended resulting in a point further along the arc in the counter clockwise direction. When the angles, $\theta_1, \dots, \theta_{i-1}$ have produced points, p_0, \dots, p_{i-1} on the arc, place the vertex of θ_i on the center of the circle and let one side of θ_i coincide with the side of the angle θ_{i-1} which is most counter clockwise, the other side of θ_i when extended, resulting in a point further along the arc A in the counterclockwise direction as shown below.



Now let $\varepsilon > 0$ be given and pick $P_1 \in \mathcal{P}(A_1)$ such that $|P_1| + \varepsilon > l(A_1)$. Then determining the angles as just described, use these angles to produce a corresponding partition of A_2 , P_2 . If $|P_2| + \varepsilon > l(A_2)$, then stop. Otherwise, pick $Q \in \mathcal{P}(A_2)$ such that $|Q| + \varepsilon > l(A_2)$ and let $P'_2 = P_2 \cup Q$. Then use the angles determined by P'_2 to obtain $P'_1 \in \mathcal{P}(A_1)$. Then $|P'_1| + \varepsilon > l(A_1)$, $|P'_2| + \varepsilon > l(A_2)$, and both P'_1 and P'_2 determine the same sequence of angles. Using Problem 36 on Page 71 about the base angles of an isosceles triangle, the two triangles are similar and so

$$\frac{|P'_1|}{|P'_2|} = \frac{R}{r}$$

Therefore

$$l(A_2) < |P'_2| + \varepsilon = \frac{r}{R} |P'_1| + \varepsilon \leq \frac{r}{R} l(A_1) + \varepsilon.$$

Since ε is arbitrary, this shows $Rl(A_2) \leq rl(A_1)$. But now reverse the argument and write

$$l(A_1) < |P'_1| + \varepsilon = \frac{R}{r} |P'_2| + \varepsilon \leq \frac{R}{r} l(A_2) + \varepsilon$$

which implies, since ε is arbitrary that $Rl(A_2) \geq rl(A_1)$ and this has proved the following theorem.

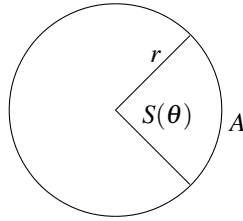
Theorem 2.3.10 *Let θ be an angle which subtends two arcs, A_R on a circle of radius R and A_r on a circle of radius r . Then denoting by $l(A)$ the length of a circular arc as described above, $Rl(A_r) = rl(A_R)$.*

Now, with this theorem, one can prove the fundamental inequality of Corollary 2.3.9.

Proof: Situate the angle θ such that one side is on the positive x axis and extend the other side till it intersects the unit circle at the point $(\cos \theta, \sin \theta)$. Then denoting the resulting arc on the circle by A , it follows that for all $P \in \mathcal{P}(A)$ the inequality $(1 - \cos \theta) + \sin \theta \geq |P| \geq \sin \theta$. It follows that $(1 - \cos \theta) + \sin \theta$ is an upper bound for all the $|P|$ where $P \in \mathcal{P}(A)$ and so $(1 - \cos \theta) + \sin \theta$ is at least as large as the sup or least upper bound of the $|P|$. This proves the top half of the inequality. The bottom half follows because $l(A) \geq L$ where L is the length of the line segment joining $(\cos \theta, \sin \theta)$ and $(1, 0)$ due to the definition of $l(A)$. However, $L \geq \sin \theta$ because L is the length of the hypotenuse of a right triangle having $\sin \theta$ as one of the sides. ■

2.3.3 The Area of a Circular Sector

Consider an arc A , of a circle of radius r which subtends an angle θ . The circular sector determined by A is obtained by joining the ends of the arc A , to the center of the circle.



The sector, $S(\theta)$ denotes the points which lie between the arc A and the two lines just mentioned. The angle between the two lines is called the central angle of the sector. The problem is to define the area of this shape. First a fundamental inequality must be obtained.

Lemma 2.3.11 *Let $1 > \varepsilon > 0$ be given. Then whenever the positive number α , is small enough,*

$$1 \leq \frac{\alpha}{\sin \alpha} \leq 1 + \varepsilon \quad (2.14)$$

and

$$1 + \varepsilon \geq \frac{\alpha}{\tan \alpha} \geq 1 - \varepsilon \quad (2.15)$$

Proof: This follows from Corollary 2.3.9 on Page 62. In this corollary, $l(A) = \alpha$ and so

$$1 - \cos \alpha + \sin \alpha \geq \alpha \geq \sin \alpha.$$

Therefore, dividing by $\sin \alpha$,

$$\frac{1 - \cos \alpha}{\sin \alpha} + 1 \geq \frac{\alpha}{\sin \alpha} \geq 1. \quad (2.16)$$

Now using the properties of the trig functions,

$$\frac{1 - \cos \alpha}{\sin \alpha} = \frac{1 - \cos^2 \alpha}{\sin \alpha (1 + \cos \alpha)} = \frac{\sin^2 \alpha}{\sin \alpha (1 + \cos \alpha)} = \frac{\sin \alpha}{1 + \cos \alpha}.$$

From the definition of the sin and cos, whenever α is small enough, $\frac{\sin \alpha}{1 + \cos \alpha} < \varepsilon$ and so 2.16 implies that for such α , 2.14 holds. To obtain 2.15, let α be small enough that 2.14 holds and multiply by $\cos \alpha$. Then for such α ,

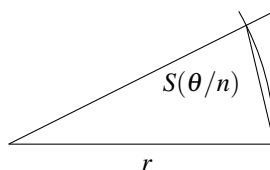
$$\cos \alpha \leq \frac{\alpha}{\tan \alpha} \leq (1 + \varepsilon) \cos \alpha$$

Taking α smaller if necessary, and noting that for all α small enough, $\cos \alpha$ is very close to 1, yields 2.15. ■

This lemma is very important in another context.

Theorem 2.3.12 *Let $S(\theta)$ denote the sector of a circle of radius r having central angle θ . Then the area of $S(\theta)$ equals $\frac{r^2}{2}\theta$.*

Proof: Let the angle which A subtends be denoted by θ and divide this sector into n equal sectors each of which has a central angle equal to θ/n . The following is a picture of one of these.



In the picture, there is a circular sector, $S(\theta/n)$ and inside this circular sector is a triangle while outside the circular sector is another triangle. Thus any reasonable definition of area would require

$$\frac{r^2}{2} \sin(\theta/n) \leq \text{area of } S(\theta/n) \leq \frac{r^2}{2} \tan(\theta/n).$$

It follows the area of the whole sector having central angle θ must satisfy the following inequality.

$$\frac{nr^2}{2} \sin(\theta/n) \leq \text{area of } S(\theta) \leq \frac{nr^2}{2} \tan(\theta/n).$$

Therefore, for all n , the area of $S(\theta)$ is trapped between the two numbers,

$$\frac{r^2}{2} \theta \frac{\sin(\theta/n)}{(\theta/n)}, \quad \frac{r^2}{2} \theta \frac{\tan(\theta/n)}{(\theta/n)}.$$

Now let $\varepsilon > 0$ be given, a small positive number less than 1, and let n be large enough that

$$1 \geq \frac{\sin(\theta/n)}{(\theta/n)} \geq \frac{1}{1 + \varepsilon}$$

and

$$\frac{1}{1 + \varepsilon} \leq \frac{\tan(\theta/n)}{(\theta/n)} \leq \frac{1}{1 - \varepsilon}.$$

Therefore,

$$\frac{r^2}{2} \theta \left(\frac{1}{1 + \varepsilon} \right) \leq \text{Area of } S(\theta) \leq \left(\frac{1}{1 - \varepsilon} \right) \frac{r^2}{2} \theta.$$

Since ε is an arbitrary small positive number, it follows the area of the sector equals $\frac{r^2}{2} \theta$ as claimed. (Why?) ■

2.4 Exercises

- Find $\cos \theta$ and $\sin \theta$ using only knowledge of angles in the first quadrant for $\theta \in \{\frac{2\pi}{3}, \frac{3\pi}{4}, \frac{5\pi}{6}, \pi, \frac{7\pi}{6}, \frac{5\pi}{4}, \frac{4\pi}{3}, \frac{3\pi}{2}, \frac{5\pi}{3}, \frac{7\pi}{4}, \frac{11\pi}{6}, 2\pi\}$.
- Prove $\cos^2 \theta = \frac{1 + \cos 2\theta}{2}$ and $\sin^2 \theta = \frac{1 - \cos 2\theta}{2}$.
- $\pi/12 = \pi/3 - \pi/4$. Therefore, from Problem 2,

$$\cos(\pi/12) = \sqrt{\frac{1 + (\sqrt{3}/2)}{2}}.$$

On the other hand,

$$\cos(\pi/12) = \cos(\pi/3 - \pi/4) = \cos \pi/3 \cos \pi/4 + \sin \pi/3 \sin \pi/4$$

and so $\cos(\pi/12) = \sqrt{2}/4 + \sqrt{6}/4$. Is there a problem here? Please explain.

- Prove $1 + \tan^2 \theta = \sec^2 \theta$ and $1 + \cot^2 \theta = \csc^2 \theta$.
- Prove that $\sin x \cos y = \frac{1}{2} (\sin(x+y) + \sin(x-y))$.
- Prove that $\sin x \sin y = \frac{1}{2} (\cos(x-y) - \cos(x+y))$.
- Prove that $\cos x \cos y = \frac{1}{2} (\cos(x+y) + \cos(x-y))$.
- Using Problem 5, find an identity for $\sin x - \sin y$.
- Suppose $\sin x = a$ where $0 < a < 1$. Find all possible values for

- | | |
|--------------|--------------|
| (a) $\tan x$ | (d) $\csc x$ |
| (b) $\cot x$ | |
| (c) $\sec x$ | (e) $\cos x$ |

- Solve the equations and give all solutions.

- | | |
|-------------------------------------|--------------------------------------|
| (a) $\sin(3x) = \frac{1}{2}$ | (e) $\sin(x+7) = \frac{\sqrt{2}}{2}$ |
| (b) $\cos(5x) = \frac{\sqrt{3}}{2}$ | (f) $\cos^2(x) = \frac{1}{2}$ |
| (c) $\tan(x) = \sqrt{3}$ | (g) $\sin^4(x) = 4$ |
| (d) $\sec(x) = 2$ | |

- Sketch a graph of $y = \sin x$.
- Sketch a graph of $y = \cos x$.
- Sketch a graph of $y = \sin 2x$.

14. Sketch a graph of $y = \tan x$.
15. Find a formula for $\sin x \cos y$ in terms of sines and cosines of $x + y$ and $x - y$.
16. Using Problem 2 graph $y = \cos^2 x$.
17. If $f(x) = A \cos \alpha x + B \sin \alpha x$, show there exists ϕ such that

$$f(x) = \sqrt{A^2 + B^2} \sin(\alpha x + \phi).$$

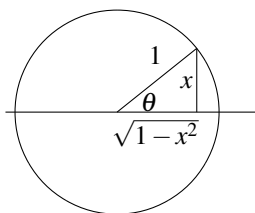
Show there also exists ψ such that $f(x) = \sqrt{A^2 + B^2} \cos(\alpha x - \psi)$. This is a very important result, enough that some of these quantities are given names. $\sqrt{A^2 + B^2}$ is called the amplitude and ϕ or ψ are called phase shifts.

18. Using Problem 17 graph $y = \sin x + \sqrt{3} \cos x$.
19. Give all solutions to $\sin x + \sqrt{3} \cos x = \sqrt{3}$. **Hint:** Use Problem 18.
20. As noted above 45° is the same angle as $\pi/4$ radians. Explain why 90° is the same angle as $\pi/2$ radians. Next find a simple formula which will change the degree measure of an angle to radian measure and radian measure into degree measure.
21. Find a formula for $\tan(\theta + \beta)$ in terms of $\tan \theta$ and $\tan \beta$.
22. Find a formula for $\tan(2\theta)$ in terms of $\tan \theta$.
23. Find a formula for $\tan\left(\frac{\theta}{2}\right)$ in terms of $\tan \theta$.
24. Show $\tan(4\theta) = \frac{4 \tan \theta - 4 \tan^3 \theta}{1 - 6 \tan^2 \theta + \tan^4 \theta}$. Use to show that

$$\frac{\pi}{4} = 4 \arctan\left(\frac{1}{5}\right) - \arctan\left(\frac{1}{239}\right).$$

Here $\arctan(x)$ is defined to be the angle in $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ whose tangent is x . That is, $\arctan(x) \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ and $\tan(\arctan(x)) = x$. This formula and others like it have been used to compute π for hundreds of years.

25. The function, \sin has domain equal to \mathbb{R} and range $[-1, 1]$. However, this function is not one to one because $\sin(x + 2\pi) = \sin x$. Show that if the domain of the function is restricted to be $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$, then \sin still maps onto $[-1, 1]$ but is now also one to one on this restricted domain. Therefore, there is an inverse function, called \arcsin which is defined by $\arcsin(x) \equiv$ the angle whose \sin is x which is in the interval, $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$. Thus $\arcsin\left(\frac{1}{2}\right)$ is the angle whose \sin is $\frac{1}{2}$ which is in $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$. This angle is $\frac{\pi}{6}$. Suppose you wanted to find $\tan(\arcsin(x))$. How would you do it? Consider the following picture which corresponds to the case where $x > 0$.



Then letting $\theta = \arcsin(x)$, the thing which is wanted is $\tan \theta$. Now from the picture, you see this is $\frac{x}{\sqrt{1-x^2}}$. If x were negative, you would have the little triangle pointing down rather than up as in the picture. The result would be the same for $\tan \theta$. Find the following:

- | | |
|------------------------|------------------------|
| (a) $\cot(\arcsin(x))$ | (c) $\csc(\arcsin(x))$ |
| (b) $\sec(\arcsin(x))$ | (d) $\cos(\arcsin(x))$ |

26. Using Problem 25 and the formulas for the trig functions of a sum of angles, find the following. Assume x, y are small and positive.

- | | |
|--------------------------|-------------------------------------|
| (a) $\cot(\arcsin(2x))$ | (d) $\cos(2\arcsin(x))$ |
| (b) $\sec(\arcsin(x+y))$ | (e) $\tan(\arcsin(x) + \arcsin(y))$ |
| (c) $\csc(\arcsin(x^2))$ | (f) $\csc(\arcsin(x) - \arcsin(y))$ |

27. The function, \cos , is onto $[-1, 1]$ but fails to be one to one. Show that if the domain of \cos is restricted to be $[0, \pi]$, then \cos is one to one on this restricted domain and still is onto $[-1, 1]$. Define $\arccos(x) \equiv$ the angle whose cosine is x which is in $[0, \pi]$. Find the following.

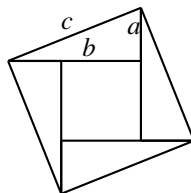
- | | |
|------------------------|------------------------|
| (a) $\tan(\arccos(x))$ | (d) $\csc(\arccos(x))$ |
| (b) $\cot(\arccos(x))$ | |
| (c) $\sin(\arccos(x))$ | (e) $\sec(\arccos(x))$ |

28. Using Problem 27 and the formulas for the trig functions of a sum of angles, find the following. Assume x, y are small and positive if desired.

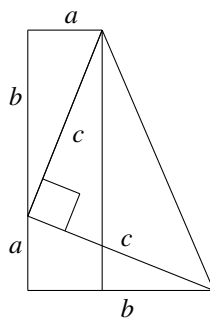
- | |
|-------------------------------------|
| (a) $\cot(\arccos(2x))$ |
| (b) $\sec(\arccos(x+y))$ |
| (c) $\csc(\arccos(x^2))$ |
| (d) $\cos(\arcsin(x) + \arccos(y))$ |
| (e) $\tan(\arcsin(x) + \arccos(y))$ |

29. The function, \arctan is defined as $\arctan(x) \equiv$ the angle whose tangent is x which is in $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$. Show this is well-defined and is the inverse function for \tan if the domain of \tan is restricted to be $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$. Find

- (a) $\cos(\arctan(x))$
 (b) $\cot(\arctan(x))$
 (c) $\sin(\arctan(x))$
 (d) $\csc(\arctan(x))$
 (e) $\sec(\arctan(x))$
30. Using the formulas for the trig functions of a sum of angles, find the following. Assume x, y are small and positive if this is helpful.
- (a) $\cot(\arctan(2x))$
 (b) $\sec(\arctan(x+y))$
 (c) $\csc(\arccos(x^2))$
 (d) $\cos(2\arctan(x) + \arcsin(y))$
31. The graphs of \tan and \cot suggest that these functions are periodic of period π verify that this is indeed the case using the identities presented.
32. Give another argument which verifies the Pythagorean theorem by supplying the details for the following argument³. Take the given right triangle and situate copies of it as shown below.



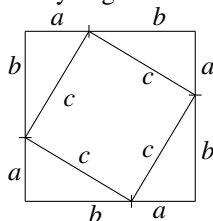
33. Another very simple and convincing proof of the Pythagorean theorem⁴ is based on writing the area of the following trapezoid two ways. Explain why the angle denoted with a square has radian measure equal to $\pi/2$ and find the area of the trapezoid two ways.



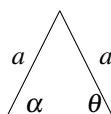
³This argument is old and was known to the Indian mathematician Bhaskar who lived 1114-1185 A.D.

⁴This argument involving the area of a trapezoid is due to James Garfield 1831-1881 who was one of the presidents of the United States. Garfield was shot early in his term as president and lingered for a couple of months during which time he was attended by a physician who did not believe in the latest knowledge about the importance of keeping wounds clean, although he was otherwise a very experienced physician who had saved the lives of many wounded men in the Civil War. It is likely that Garfield would have survived if he had received better medical care. They never found the bullet and kept probing the wound looking for it, thus introducing more infection. If you look up Garfield, you will find many other interesting things. He was made the republican nominee by acclamation.

34. Make up your own proof of the Pythagorean theorem based on the following picture.



35. If A, B are two angles in a triangle, and $\cos(A) = \cos(B)$, explain why $A = B$. **Hint:** Note that \cos is one to one on $[0, \pi]$ and all angles of a triangle have radian measure less than π .
36. An isosceles triangle is one which has two equal sides. For example the following picture is of an isosceles triangle



the two equal sides having length a . Show the “base angles” θ and α are equal. **Hint:** You might want to use the law of cosines and the observation in the above problem.

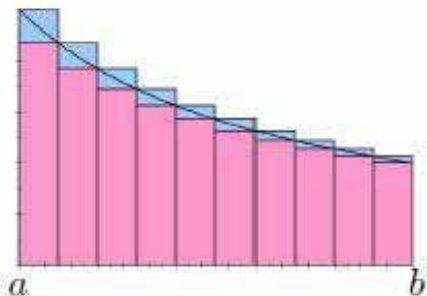
2.5 Exponential and Logarithmic Functions

Logarithms were first considered by Napier⁵ and were based on an assumption that an exponential function exists. However, by the latter half of the seventeenth century, it was realized that the best way to consider exponential functions and logarithms is through the general approach discussed in this section where the natural logarithm \ln is defined first and used to obtain the exponential function.

Let $A(a, b)$ denote the area of the region $R(a, b)$ which is under the graph of $y = 1/x$, above the x axis, and between the graphs of the lines $x = a$ and $x = b$.

⁵Napier was a Scottish nobleman. He lived from 1550 - 1617. In addition to inventing logarithms, he also predicted the end of the world would occur in 1770 based on his study of the Book of Revelations and other ancient manuscripts. Those who developed calculus were often interested in theology and some made strange predictions determining the end of the world. A belief in the inerrancy of the Bible can lead to many strange conclusions. Napier was certainly not unique in this. We don't hear such predictions all that often now, but they were very fashionable in the eighteenth and nineteenth centuries. Napier did not worry about the sort of thing in this section.

Upper and Lower sums for $y = 1/x$



The above picture illustrates lower and upper sums for this region. The sum of the areas of the rectangles below the graph is a lower sum and the sum of the rectangles which are each too tall is an upper sum.

The area is defined to be the number which is between all such upper sums and all such lower sums. Denote by $U_n(a, b)$ the sum of the areas of the rectangles which enclose $R(a, b)$ in case there are n of them having equal length and $L_n(a, b)$ the sum of the areas of the rectangles which are contained in $R(a, b)$. You would think that the approximation to the area would improve by having more of these rectangles.

In the picture $n = 10$. What is $U_n(a, b) - L_n(a, b)$, the discrepancy between the two sums of areas of rectangles? The width of each rectangle is $(b - a)/n$. Thus $U_n(a, b) =$

$$\begin{aligned} & a^{-1} \frac{b-a}{n} + \left(a + \frac{b-a}{n}\right)^{-1} \frac{b-a}{n} + \cdots + \left(a + (n-1) \frac{b-a}{n}\right)^{-1} \frac{b-a}{n} \\ &= \sum_{k=0}^{n-1} \left(a + k \frac{b-a}{n}\right)^{-1} \frac{b-a}{n} \end{aligned}$$

a similar formula holding for $L_n(a, b)$ in which $\sum_{k=0}^{n-1}$ is replaced with $\sum_{k=1}^n$. Hence,

$$\begin{aligned} & U_n(a, b) - L_n(a, b) = \\ & \sum_{k=0}^{n-1} \left(a + k \frac{b-a}{n}\right)^{-1} \frac{b-a}{n} - \sum_{k=1}^n \left(a + k \frac{b-a}{n}\right)^{-1} \frac{b-a}{n} \\ &= \left(a^{-1} - (a + b - a)^{-1}\right) \frac{b-a}{n} = \frac{1}{ab} (b-a)^2 \frac{1}{n} \end{aligned}$$

This shows that if n is large, the sum of the areas of the small rectangles inside the region and the sum of the areas of the large rectangles enclosing the region, are both approximately equal to what should be defined as the area of the region just described.

Now notice that if $r > 0$,

$$U_n(ar, br) = U_n(a, b), \quad L_n(ar, br) = L_n(a, b).$$

The reason for this is that the width of the rectangles in the sum for $U_n(ar, br)$ is multiplied by r and the height is multiplied by $1/r$, leaving the area unchanged for each rectangle in the sum. Similar considerations apply to $L_n(a, b)$. Therefore,

$$\begin{aligned} A(a, b) - A(ra, rb) &\leq \overset{\text{too large}}{U_n(a, b)} - \overset{\text{too small}}{L_n(ra, rb)} \\ &= U_n(a, b) - L_n(a, b) \leq \frac{1}{ab} (b-a)^2 \frac{1}{n} \end{aligned}$$

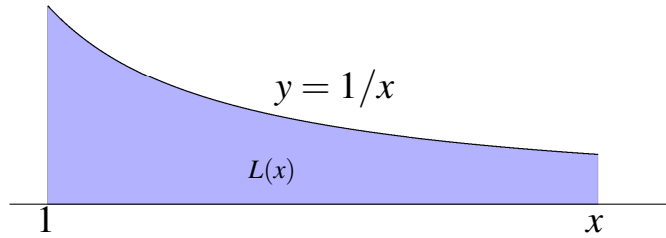
Similarly,

$$\begin{aligned} A(ra, rb) - A(a, b) &\leq U_n(ra, rb) - L_n(a, b) \\ &= U_n(a, b) - L_n(a, b) \leq \frac{1}{ab} (b-a)^2 \frac{1}{n} \end{aligned}$$

Thus $|A(ra, rb) - A(a, b)| \leq \frac{1}{ab} (b-a)^2 \frac{1}{n}$ and since n is arbitrary, $A(ra, rb) = A(a, b)$. This also holds for $a > b$ if $A(a, b) \equiv -A(b, a)$. This is summarized in the following lemma.

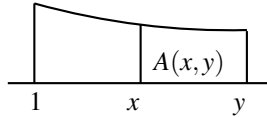
Lemma 2.5.1 *Let $a, b > 0$ and denote by $A(a, b)$ the area of the region which is bounded by the lines $x = a, x = b$, the graph of $y = 1/x$, and the x axis if $a \leq b$ and if $a > b$, then $A(a, b) \equiv -A(b, a)$. Then if $r > 0, A(ra, rb) = A(a, b)$.*

Definition 2.5.2 *For $a > b, A(a, b) \equiv -A(b, a)$. For $x > 0$ define $L(x)$ as follows. Letting $A(a, b)$ be as just defined, $L(x) \equiv A(1, x)$.*



Lemma 2.5.3 *Whenever x, y are positive, $A(x, y) = A(1, y) - A(1, x)$. Whenever x, y positive, $L\left(\frac{x}{y}\right) = L(x) - L(y)$ and $L(xy) = L(x) + L(y)$, $L(1) = 0$, and $L(y) = -L(y^{-1})$.*

Proof: First suppose $x \leq y$. If $1 \leq x$, the claim is clearly so from the definition of $A(a, b)$ as area under the curve for $a \leq x \leq b$.



If $x < 1$, Then similarly, $A(x, y) = A(x, 1) + A(1, y) = A(1, y) - A(1, x)$. If $y \leq 1$, then $A(x, y) + A(y, 1) = A(x, 1) = -A(1, x)$ and so $A(x, y) = -A(y, 1) - A(1, x) = A(1, y) - A(1, x)$ again. This completes the case that $x \leq y$.

Now suppose $x > y$. Then from the definition and what was just shown, $A(x, y) \equiv -A(y, x) = -(A(1, x) - A(1, y)) = A(1, y) - A(1, x)$.

For the claim about L , and Lemma 2.5.1, $L\left(\frac{x}{y}\right) \equiv A\left(1, \frac{x}{y}\right) = A\left(\frac{y}{y}, \frac{x}{y}\right) = A(y, x) = A(1, x) - A(1, y) = L(x) - L(y)$. Finally, from the above,

$$L(xy) = L\left(\frac{x}{y^{-1}}\right) = L(x) - L(y^{-1})$$

Now $L(y^{-1}) \equiv A(1, y^{-1}) = A\left(\frac{y}{y}, \frac{1}{y}\right) = A(y, 1) = -A(1, y) = -L(y)$ and so the above yields $L(xy) = L(x) + L(y)$. In particular, $L(1) = L(1^2) = L(1) + L(1)$ so $L(1) = 0$ and $L(y) = -L(y^{-1})$. ■

Theorem 2.5.4 *The function of Definition 2.5.2 for $x, y > 0$ has the following properties.*

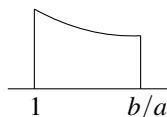
1. $L(xy) = L(x) + L(y)$, $L(1) = 0$, $L\left(\frac{x}{y}\right) = L(x) - L(y)$, $L(x) = -L\left(\frac{1}{x}\right)$
2. $x \rightarrow L(x)$ is strictly increasing so L is one to one and $L(x) = -L\left(\frac{1}{x}\right) < 0$ if $x < 1$ while $L(x) > 0$ if $x > 1$.
3. If $x > 0$, then if $y > x/2$, $|L(y) - L(x)| \leq \frac{2|x-y|}{|x|}$.
4. $L: (0, \infty) \rightarrow \mathbb{R}$ is onto.

Proof: The first claim is in the above lemma.

Consider the second claim. To see this, first note that if $x < y$, then $L(y) - L(x) = L(y/x)$ by the first claim, and $L(y/x) > 0$ so it follows that L is strictly increasing. Hence the function L is also one to one. From the definition and the first part, $L(x) = -L\left(\frac{1}{x}\right) < 0$ if $x < 1$ and $L(x) > 0$ if $x > 1$.

Now consider the third claim. If $a \leq b$, then $L(b/a) + L(a) = L(b)$ and so, from the definition in terms of area,

$$0 \leq L(b) - L(a) = L(b/a) = A(1, b/a) \leq \left(\frac{b}{a} - 1\right) 1 = \frac{b-a}{a}.$$



If $x/2 < y \leq x$. Then, from the above,

$$|L(x) - L(y)| = L(x) - L(y) \leq \frac{x-y}{y} \leq \frac{|x-y|}{x/2} = \frac{2|x-y|}{x}$$

If $y > x$, then since $2/x > 1/x$,

$$|L(x) - L(y)| = L(y) - L(x) \leq \frac{y-x}{x} \leq \frac{2|x-y|}{x}$$

Finally consider the last claim. From the definition, it follows that $L(2) > 1/2$. (Draw a picture.) Therefore, from the first claim,

$$L(2^n) = L\left(\overbrace{2 \times 2 \times \cdots \times 2}^{n \text{ times}}\right) = nL(2) > n/2, \quad L\left(\frac{1}{2^n}\right) = nL(1/2) < -n/2.$$

Thus $L(x)$ achieves arbitrarily large and arbitrarily small values for positive x . Let $y \in \mathbb{R}$ and let $S \equiv \{x : L(x) > y\}$. Then $S \neq \emptyset$. Let $l = \inf(S)$. It follows that $l > 0$ because L achieves values smaller than y on positive numbers. Then there exists $x_n \in [l, l+n^{-1}) \cap S$ since otherwise l is not the greatest lower bound. ($l+n^{-1}$ would be a larger lower bound.)

Also let $y_n \in (l - 1/n, l)$ for n large enough that $l - 1/n > 0$. Is $L(l) = y$? For large enough n , x_n, y_n are both larger than $l/2$ so from Claim 3,

$$\begin{aligned} |y - L(l)| &< L(x_n) - L(y_n) = (L(x_n) - L(l)) + (L(l) - L(y_n)) \\ &\leq \frac{2|x_n - l|}{l} + \frac{2|y_n - l|}{l} < \frac{2}{l} \frac{2}{n} \end{aligned}$$

Since n is arbitrarily large, $y = L(l)$. Therefore, L is onto \mathbb{R} . ■

This function L will be denoted as \ln . It is the natural logarithm.

Definition 2.5.5 *Since \ln is one to one and onto, there exists a real number e such that $\ln(e) = 1$. This number e is called Euler's number.*

It can be computed directly from the above definition of \ln . You would get a good table of values of \ln using the above definition and then go backwards in the table to obtain an approximation for e . There are of course much more sophisticated ways to find it, and here it is to several decimal places. $e = 2.7183$

From knowledge of \ln and its properties, it is easy to get the existence of an exponential function. Let $\exp : \mathbb{R} \mapsto (0, \infty)$ be the inverse of \ln .

$$\exp \text{ is defined on all of } \mathbb{R} \quad (2.17)$$

$$\exp(x+y) = \exp(x) \exp(y) \quad (2.18)$$

$$\exp : \mathbb{R} \mapsto (0, \infty) \text{ is one to one and onto} \quad (2.19)$$

$$\exp(0) = 1 \quad (2.20)$$

Each is satisfied. Since \ln maps $(0, \infty)$ onto \mathbb{R} , its inverse function is defined on \mathbb{R} and has values which are positive numbers. This inverse function \exp is one to one and onto. Indeed, if $\exp(y_1) = \exp(y_2)$, then since each is positive, you can take \ln of both sides and conclude that $y_1 = y_2$. Thus \exp is one to one. If $y \in (0, \infty)$, then $\ln(y) \in \mathbb{R}$ and so $y = \exp(\ln(y))$. Thus \exp is also onto.

$$\begin{aligned} \ln(\exp(x+y)) &= x+y \\ \ln(\exp(x)\exp(y)) &= \ln(\exp(x)) + \ln(\exp(y)) = x+y \end{aligned}$$

Since \ln is one to one, this verifies 2.18.

Observation 2.5.6 *From 2.18, it follows that for n an integer, $(\exp(x))^n = \exp(nx)$. Also, for m an integer, it follows from Theorem 2.5.4 that for n an integer, $n \ln(x) = \ln(x^n)$. If n is a positive integer, this is obvious from that theorem so consider $-n$ with n positive. By the same theorem,*

$$-n \ln(x) = n \ln(1/x) = \ln(1/x^n) = \ln(x^{-n}).$$

2.6 The Function b^x

You have no idea what $2^{\sqrt{2}}$ is. You do know what 2^n is for n an integer. You also know what $2^{m/n}$ is for m, n integers. It is $\sqrt[n]{2^m}$. For b a positive real number and r a real number, define $b^r \equiv \exp(r \ln b)$. Does this definition contradict what we already know?

Proposition 2.6.1 For $b > 0$ and r a real number, define $b^r \equiv \exp(r \ln(b))$. Then if $r = m/n$, for m, n integers, $\exp(\frac{m}{n} \ln b) = \sqrt[n]{b^m}$, the positive n^{th} root of b^m .

Proof: From the above observation,

$$\begin{aligned} \left(\exp\left(\frac{m}{n} \ln b\right) \right)^n &= \exp\left(n \left(\frac{m}{n} \ln b\right)\right) \\ &= \exp(m \ln(b)) = \exp(\ln(b^m)) = b^m \end{aligned}$$

Therefore, taking n^{th} roots, $\exp(\frac{m}{n} \ln b) = \sqrt[n]{b^m}$. Recall that from Theorem 1.10.2 there is a unique positive n^{th} root of a positive number so everything makes sense here. ■

This is a very important observation because it shows that if we define b^r as the expression involving known functions $\exp(r \ln(b))$, there is no contradiction between this definition and what was already accepted for r rational. It is not like what is done in some religions where new policies contradict that which was earlier identified as god's will, and everyone is supposed to choose to believe both even though they contradict. Here we can make the following definition of b^r for r real and $b > 0$ with no cognitive dissonance. This definition avoids the pretense that we know what a number raised to a real power means when we really don't.

Definition 2.6.2 Let $b > 0$ and let r be a real number. Then $b^r \equiv \exp(r \ln(b))$.

Proposition 2.6.3 The usual rules of exponents hold.

Proof: These properties follow directly from the definition.

$$b^{r+\hat{r}} \equiv \exp((r+\hat{r}) \ln(b)) = \exp(r \ln(b)) \exp(\hat{r} \ln(b)) = b^r b^{\hat{r}}.$$

$$b^0 \equiv \exp(0 \ln(b)) = \exp(0) = 1.$$

$$(b^r)^{\hat{r}} \equiv \exp(\hat{r} \ln(b^r)) \equiv \exp(\hat{r} \ln(\exp(r \ln(b)))) = \exp(\hat{r} r \ln(b)) \equiv b^{r\hat{r}} \quad \blacksquare$$

Observation 2.6.4 Note that for e the number such that $\ln(e) = 1$, $e^x \equiv \exp(x \ln(e)) = \exp(x)$ and so from now on, I will use either e^x or $\exp(x)$ because they are exactly the same thing.

If you have $x = f(t)$ and $y = g(t)$ for t in some interval, then $(f(t), g(t))$ for $t \in [a, b]$ traces out a curve in the plane. These equations $x = f(t)$ and $y = g(t)$ are called parametrizations of this curve. This will be discussed more later but it is convenient to introduce the term now.

Definition 2.6.5 The hyperbolic functions, denoted as $\cosh(x)$, $\sinh(x)$ are defined as follows:

$$\cosh(x) \equiv \frac{e^x + e^{-x}}{2}, \quad \sinh(x) \equiv \frac{e^x - e^{-x}}{2}$$

The reason these are called hyperbolic functions is that

$$\cosh^2(x) - \sinh^2(x) = 1$$

Thus if $x = \cosh(t)$ and $y = \sinh(t)$, then $x^2 - y^2 = 1$ so $(x, y) = (\cosh(t), \sinh(t))$ parametrizes a hyperbola given by $x^2 - y^2 = 1$. The circular functions $\cos(t)$, $\sin(t)$ are so called because $\cos^2(t) + \sin^2(t) = 1$. Thus if $(x, y) = (\cos(t), \sin(t))$, this parametrizes a circle.

Note that $\cosh(t) + \sinh(t) = e^t$ and $\cosh(t) = \cosh(-t)$ while $\sinh(-t) = -\sinh(t)$. These are the even and odd parts of the function $t \rightarrow e^t$. This has just given all essential features of the hyperbolic functions.

2.7 Applications

2.7.1 Interest Compounded Continuously

It is possible to compound interest continuously. If time is measured in years and if the interest rate is r per year, compounded n times a year, then it can be shown that the amount after t years is $P\left(1 + \frac{r}{n}\right)^{nt}$. The idea is to let n get larger and larger. From material presented later, the amount becomes increasingly close to $P\exp(rt)$. This explains the following procedure for compounding interest continuously.

Procedure 2.7.1 *If the interest rate is r and the interest is compounded continuously, to find the future value after t years you compute $P\exp(rt)$.*

Example 2.7.2 *The interest rate is 10% and the payment is \$1000. What is the future value after 5 years if interest is compounded daily and if interest is compounded continuously.*

To compound daily, you would have 365 payment periods in each year so the future value is $1000\left(1 + \frac{.1}{365}\right)^{5 \times 365} = \1648.60 . Now compounding it continuously you get $1000\exp(5 \times .1) = \1648.70 . You see, compounding continuously is better than compounding daily. If you wait 5 years you get an extra 10 cents. Well, every little bit helps.

2.7.2 Exponential Growth and Decay

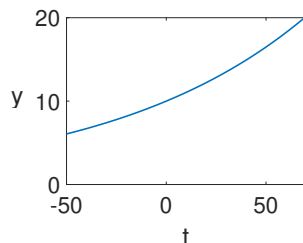
Suppose you have a bacteria culture and you feed it all it needs and there is no restriction on its growth due to crowding for example. Then in this case, the rate of growth is proportional to the amount of bacteria present. This is because the more you have, the more bacteria there are to divide and make new bacteria. Consider equally spaced intervals of time such that n of them equal one unit of time where n is large. The unit might be years, days, etc. Also let A_k denote the amount of whatever is growing at the end of the k^{th} increment of time. In exponential growth

$$A_{k+1} - A_k \approx rA_k(1/n), A_0 = P$$

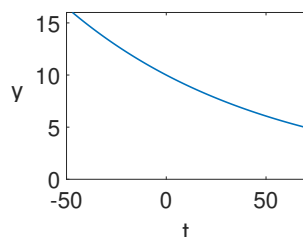
where r is a proportionality constant called the growth rate and the above difference equation should give a good description of the amount provided n is large enough. The symbol \approx indicates approximately equal. Now this is easy to solve. $A_{k+1} \approx \left(1 + \frac{r}{n}\right)A_k$, $A_0 = P$ and you look for $A_k = \alpha^k$ and find α . This is easily seen to be $\alpha = \left(1 + \frac{r}{n}\right)$ and so $A_k \approx P\left(1 + \frac{r}{n}\right)^k$. Now let $A(t)$ denote the amount at time t . What is it? There are tn time intervals so k goes up to tn and you get

$$A(t) \approx P\left(1 + \frac{r}{n}\right)^{nt} \approx P\left(1 + \frac{rt}{nt}\right)^{nt}$$

the approximation for $A(t)$ getting better as $n \rightarrow \infty$. As will be shown later, this becomes very close to $A(t) = P\exp(rt)$ when n is large which is the formula for exponential growth. As an example, consider $P = 10$ and $r = .01$. Then the following is the graph of the function $y = 10e^{.01t}$.



When the rate of change is negative, the process is called exponential decay. This is the process which governs radioactive substances. It is the same formula which results, only this time it is of the form $A(t) = P \exp(-rt)$ where $r > 0$. Consider the same example only this time consider $y = 10e^{-.01t}$.



Exercise 2.7.3 Carbon 14 has a half life of 5730 years. This means that if you start with a given amount of it and wait 5730 years, there will be half as much left. Carbon 14 is assumed to be constantly created by cosmic rays hitting the atmosphere so that the proportion of carbon in a living organism is the same now as it was a long time ago. This is of course an assumption and there is evidence it is not true but this does not concern us here. When the living thing dies, it quits replenishing the carbon 14 and so that which it has decays according to the above half life. By measuring the amount in the remains of the dead thing and comparing with what it had when it was alive, one can determine an estimate for how long it has been dead. Suppose then you measure the amount of carbon 14 in some dead wood and find there is $1/3$ the amount there would have been when it was alive. How long ago did the tree from which the wood came die?

Let $A(t)$ be the amount of carbon 14 in the sample and let A_0 be the amount when it died. Then $A(t) = A_0 \exp(-rt)$. By assumption $.33A_0 = A_0 \exp(-rt)$ and cancelling the A_0 one can solve for t as follows. $\ln(.33) = -rt$. If I knew what r was, I could then solve for t . The half life is 5730 and so $.5 = \exp(-r5730)$ and so $\ln(.5) = -r(5730)$ from properties of \ln described above, $-\ln(1/2) = \ln((1/2)^{-1}) = \ln(2)$. Therefore, $r = \frac{\ln 2}{5730} = 1.2097 \times 10^{-4}$. To get this number, I just used the computer. As mentioned above \ln has been tabulated. Therefore, in the problem of interest, $\ln(.33) = -(1.2097 \times 10^{-4})t$ and so

$$t = \frac{\ln(.33)}{-1.2097 \times 10^{-4}} = 9164.8 \text{ years.}$$

So how did they find the half life of carbon 14? Did Noah have a sample of recently dead wood in the ark and make some measurements which he recorded in the Book of the

Law of Noah which were then compared to measurements made in the twentieth century using chronology determined by Bishop Ussher to determine that exactly 5730 years had passed? Actually, this is not the way it was done. The half life was also not established by the decree of omniscient scientists. When I was young, I was constantly asked to believe what “scientists” thought but never given any reason why they thought what they did. We don’t have to do this in math.

Example 2.7.4 Find the half life of a radioactive substance if after 5 years there is .999 395 of the original amount present.

This says $.999\,395 = \exp(-5r)$ and so $r = \frac{\ln(.999\,395)}{-5} = 1.210\,366\,17 \times 10^{-4}$. Now to find the half life T , you need to solve the equation $\frac{1}{2} = e^{-(1.210\,366\,17 \times 10^{-4})T}$. Thus $\ln(.5) = -(1.210\,366\,17 \times 10^{-4})T$ and so $T = \frac{\ln(.5)}{-(1.210\,366\,17 \times 10^{-4})} = 5726$. Using the known properties of exponential decay, you can compute the half life without waiting for over 5000 years.

2.7.3 The Logistic Equation

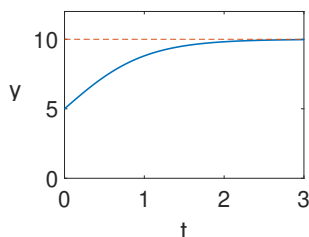
There is a class of functions called logistic functions. These are studied in differential equations and the derivation of these functions follows from the techniques of this subject. Roughly you have a population called y and the rate at which it grows is proportional to itself and $(1 - y/K)$, the constant of proportionality being r . The idea is that the growth is exponential which is attenuated by an approach to the maximum population K . When the differential equation (discussed later) is solved, it yields a logistic function,

$$y(t) = \frac{K}{1 + e^{-rt}CK}$$

Typically t measures time and y is the population. One interesting application is to the population of cancer cells or the size of a tumor. Typically $r > 0$ and $K, C > 0$. The constant C is computed from information on y when $t = 0$. The constants r, K are given parameters. Typically the initial value of y is given to be less than K and K is sometimes called the carrying capacity. In population models, it is the maximum population. Here is an example.

Example 2.7.5 Suppose $r = 2, K = 10$, and $y(0) = 5$. Find C and graph the resulting function y .

From the information, you need to have $5 = \frac{10}{1+10C}$. Therefore, $C = 1/10$ and the function is of the form $y(t) = \frac{10}{1+\exp(-2t)}$. Now here is the graph of this function.



The red line is the asymptote $y = 10$. Note how as t increases, the function becomes increasingly close to the value 10. This illustrates how 10 is the largest possible value for y , the maximum population.

2.8 Using MATLAB to Graph

Suppose you want to graph the function $y = \sin(x^2)$ for $x \in [0, 5]$. In MATLAB, you would do the following:

```
x=[0:.01:4];
plot(x,sin(x.^2),'LineWidth',2)
```

Then press enter and it will produce the graph of this function. Note that x is a list of numbers between 0 and 4 which are .01 apart. $x.^2$ says to make a list of numbers obtained by squaring each number in the original list. This is why you need $.$ rather than simply $^$. You also need to press shift enter to get to a new line. Don't forget to put ; after the first line. You don't want to see the list of numbers.

2.9 Exercises

1. Define logarithms to the base b for b a positive real number, $b \neq 1$, as follows. For $x > 0$

$$\log_b(x) \equiv \frac{\ln(x)}{\ln(b)}$$

Show \log_b is one to one and maps onto \mathbb{R} . Then show it satisfies the same properties as \ln . That is,

$$\log_b(xy) = \log_b(x) + \log_b(y)$$

Also show that $b^{\log_b(x)} = x$ and $\log_b(b^x) = x$ and $\log_b(a^x) = x \log_b(a)$ whenever a is a positive real number.

2. Show that $\log_e(x) = \ln(x)$.
3. Solve the following equation for x : $\log_4(2^x) + 3 \log_3(9^x) = 2$
4. Show that for positive a and x , $\log_a(x) = \frac{\log_b(x)}{\log_b(a)}$
5. Simplify $\log_b(a) \log_a(b)$ where a, b are positive numbers not equal to 1.
6. Solve the following equations in terms of logarithms. **Hint:** Take natural logarithms of both sides.

(a) $2^{3x+1} = 3^{2x-2}$.

(b) $\frac{5^{x-1}}{2^{3x+1}} = 7^x$

(c) $5^x 7^{x+1} = 2^x$

7. Find x such that $\log_x(8) = 3$.
8. Find x such that $\log_x\left(\frac{1}{16}\right) = 4$.

9. If $1 < a < b$ and $x > 1$, how are $\log_a(x)$ and $\log_b(x)$ related? Which is larger? Explain why.
10. Find without using a calculator $\log_3(27), \log_2(64), \log_{10}(1000), \log_{1/2}(8)$.
11. Find the domain of the function of x given by

$$\log_3 \left(\frac{x+1}{(x-1)(x+2)} \right)$$

Hint: You need x to be such that the expression inside the parenthesis is positive and makes sense. Thus you can't have for example $x = 1$.

12. Find the domain of the function $f(x) = \sqrt{\ln \left(\frac{x+1}{x+2} \right)}$.
13. Find all solutions to $\log_2(x+4) = \log_4(x+16)$.
14. If the interest rate is 4% compounded continuously, how long does it take a given amount of money to double? This means the rate per year is .04.
15. If the interest rate is 6% compounded continuously, how long does it take a given amount of money to double? This means the rate per year is .06.
16. The population of bacteria grows exponentially. It is observed that every hour this population doubles. How long will it take to have eight times as many bacteria as at the beginning?
17. A pesticide has a half life of 27 years. How long will it take to have only 1/4 the initial amount?
18. Suppose 5% interest is compounded continuously and you make a payment of \$100 at the end of every year, starting with an initial \$1000. How much will you have at the end of 10 years?
19. Measurements are taken of an exponentially decaying substance and it is found that after 5 years there is .9 of the amount which was present at the start. What is the half life of this substance?
20. Consider the logistic equation

$$y = \frac{K}{1 + e^{-rt}CK}$$

where $y(0) = 20$ and $K = 30, r = 1$. Find C in the logistic equation to conform with this information. Then graph the resulting function.

21. In the logistic equation, explain why for large t it will always be close to K . In other words, explain why it has a horizontal asymptote of the form $y = K$.
22. Suppose $f: \mathbb{R} \rightarrow \mathbb{R}$ is any function. Let

$$h(t) \equiv (f(t) + f(-t)) \frac{1}{2}, g(t) = \frac{1}{2} (f(t) - f(-t)).$$

Show that h is even, meaning $h(t) = h(-t)$, g is odd meaning $g(-t) = -g(t)$ and that $f = g + h$. Then h is called the even part of f and g is called the odd part of f . The hyperbolic functions are defined this way as the even and odd parts of the exponential function.

23. Solve for x . $3^{\log_{27}(x)} = 4$.

2.10 Videos

[circular functions](#) [logs and exponentials](#)

Chapter 3

Sequences and Compactness

This chapter is devoted to the fundamental properties of the real line which make all existence theorems in Calculus possible. Of course you can follow stupid algorithms without these things, but if you wish to understand what is going on, you need the concepts of this chapter.

3.1 Sequences

Functions defined on the set of integers larger than a given integer are called sequences. This turns out to be somewhat easier to consider in terms of limits than functions defined on \mathbb{R} which is why I am placing this early.

Definition 3.1.1 *A function whose domain is defined as a set of the form*

$$\{k, k+1, k+2, \dots\}$$

for k an integer is known as a sequence. Thus you can consider

$$f(k), f(k+1), f(k+2),$$

etc. Usually the domain of the sequence is either \mathbb{N} , the natural numbers consisting of $\{1, 2, 3, \dots\}$ or the nonnegative integers, $\{0, 1, 2, 3, \dots\}$. Also, it is traditional to write f_1, f_2 , etc. instead of $f(1), f(2), f(3)$ etc. when referring to sequences. In the above context, f_k is called the first term, f_{k+1} the second and so forth. It is also common to write the sequence, not as f but as $\{f_i\}_{i=k}^{\infty}$ or just $\{f_i\}$ for short.

Example 3.1.2 *Let $\{a_k\}_{k=1}^{\infty}$ be defined by $a_k \equiv k^2 + 1$.*

This gives a sequence. In fact, $a_7 = a(7) = 7^2 + 1 = 50$ just from using the formula for the k^{th} term of the sequence.

It is nice when sequences come in this way from a formula for the k^{th} term. However, this is often not the case. Sometimes sequences are defined recursively. This happens, when the first several terms of the sequence are given and then a rule is specified which determines a_{n+1} from knowledge of a_1, \dots, a_n . This rule which specifies a_{n+1} from knowledge of a_k for $k \leq n$ is known as a recurrence relation.

Example 3.1.3 Let $a_1 = 1$ and $a_2 = 1$. Assuming a_1, \dots, a_{n+1} are known, $a_{n+2} \equiv a_n + a_{n+1}$.

Thus the first several terms of this sequence, listed in order, are 1, 1, 2, 3, 5, 8, \dots . This particular sequence is called the Fibonacci sequence and is important in the study of reproducing rabbits. Note this defines a function without giving a formula for it. Such sequences occur naturally in the solution of differential equations using power series methods and in many other situations of great importance.

3.2 Exercises

- Let $g(t) \equiv \sqrt{2-t}$ and let $f(t) = \frac{1}{t}$. Find $g \circ f$. Include the domain of $g \circ f$.
- Give the domains of the following functions.

(a) $f(x) = \frac{x+3}{3x-2}$ (b) $f(x) = \sqrt{x^2-4}$ (c) $f(x) = \sqrt{4-x^2}$	(d) $f(x) = \sqrt{\frac{x-4}{3x+5}}$ (e) $f(x) = \sqrt{\frac{x^2-4}{x+1}}$
---	---
- Let $f: \mathbb{R} \rightarrow \mathbb{R}$ be defined by $f(t) \equiv t^3 + 1$. Is f one to one? Can you find a formula for f^{-1} ?
- Suppose $a_1 = 1, a_2 = 3$, and $a_3 = -1$. Suppose also that for $n \geq 4$ it is known that $a_n = a_{n-1} + 2a_{n-2} + 3a_{n-3}$. Find a_7 . Are you able to guess a formula for the k^{th} term of this sequence?
- Let $f: \{t \in \mathbb{R} : t \neq -1\} \rightarrow \mathbb{R}$ be defined by $f(t) \equiv \frac{t}{t+1}$. Find f^{-1} if possible.
- A function $f: \mathbb{R} \rightarrow \mathbb{R}$ is a strictly increasing function if whenever $x < y$, it follows that $f(x) < f(y)$. If f is a strictly increasing function, does f^{-1} always exist? Explain your answer.
- Let $f(t)$ be defined by

$$f(t) = \begin{cases} 2t+1 & \text{if } t \leq 1 \\ t & \text{if } t > 1 \end{cases}.$$
 Find f^{-1} if possible.
- Suppose $f: D(f) \rightarrow R(f)$ is one to one, $R(f) \subseteq D(g)$, and $g: D(g) \rightarrow R(g)$ is one to one. Does it follow that $g \circ f$ is one to one?
- If $f: \mathbb{R} \rightarrow \mathbb{R}$ and $g: \mathbb{R} \rightarrow \mathbb{R}$ are two one to one functions, which of the following are necessarily one to one on their domains? Explain why or why not by giving a proof or an example.

(a) $f+g$ (b) fg	(c) f^3 (d) f/g
-----------------------	------------------------
- Draw the graph of the function $f(x) = x^3 + 1$.

11. Draw the graph of the function $f(x) = x^2 + 2x + 2$.
12. Draw the graph of the function $f(x) = \frac{x}{1+x}$.
13. Suppose $a_n = \frac{1}{n}$ and let $n_k = 2^k$. Find b_k where $b_k = a_{n_k}$.
14. Suppose $f(x) + f\left(\frac{1}{x}\right) = 7x$ and f is a function defined on $\mathbb{R} \setminus \{0\}$, the nonzero real numbers. Find all values of x where $f(x) = 1$ if there are any. Does there exist any such function?
15. Does there exist a function f , satisfying $f(x) - f\left(\frac{1}{x}\right) = 3x$ which has both x and $\frac{1}{x}$ in the domain of f ?
16. In the situation of the Fibonacci sequence show that the formula for the n^{th} term can be found and is given by

$$a_n = \frac{\sqrt{5}}{5} \left(\frac{1+\sqrt{5}}{2} \right)^n - \frac{\sqrt{5}}{5} \left(\frac{1-\sqrt{5}}{2} \right)^n.$$

Hint: You might be able to do this by induction but a better way would be to look for a solution to the recurrence relation, $a_{n+2} \equiv a_n + a_{n+1}$ of the form r^n . You will be able to show that there are two values of r which work, one of which is $r = \frac{1+\sqrt{5}}{2}$. Next you can observe that if r_1^n and r_2^n both satisfy the recurrence relation then so does $cr_1^n + dr_2^n$ for any choice of constants c, d . Then you try to pick c and d such that the conditions, $a_1 = 1$ and $a_2 = 1$ both hold. This general approach often works for finding solutions to such recurrence relations.

17. In an annuity, you make constant payments P at the end of each payment period. These accrue interest at the rate of r per payment period. Let A_n be the amount at the end of the n^{th} payment period. Then $rA_n + A_n + P = (1+r)A_n + P = A_{n+1}$ and the initial amount is $0 = A_0, P = A_1$. Find A_n . **Hint:** Look for $A_n = Cz^n + s$. $A_0 = 0$ so
18. A well known puzzle consists of three pegs and several disks each of a different diameter, each having a hole in the center which allows it to be slid down each of the pegs. These disks are piled one on top of the other on one of the pegs, in order of decreasing diameter, the larger disks always being below the smaller disks. The problem is to move the whole pile of disks to another peg such that you never place a disk on a smaller disk. If you have n disks, how many moves will it take? Of course this depends on n . If $n = 1$, you can do it in one move. If $n = 2$, you would need 3. Let A_n be the number required for n disks. Then in solving the puzzle, you must first obtain the top $n - 1$ disks arranged in order on another peg before you can move the bottom disk of the original pile. This takes A_{n-1} moves. Explain why $A_n = 2A_{n-1} + 1, A_1 = 1$ and give a formula for A_n . Look for one in the form $A_n = Cr^n + s$. This puzzle is called the Tower of Hanoi. When you have found a formula for A_n , explain why it is not possible to do this puzzle if n is very large.

3.3 The Limit of a Sequence

A little later, the limit of functions of real variables will be important. A sequence is just a special kind of function and it turns out that it is easier to consider the limit of a sequence.

This also helps considerably in understanding certain other concepts like continuity of a function also presented later. This is why I am including this topic early in the book, to make more difficult concepts easier to understand.

The concept of the limit of a sequence was defined precisely by Bolzano.¹ It is now expressed as follows.

Definition 3.3.1 A sequence $\{a_n\}_{n=1}^{\infty}$ converges to a , written

$$\lim_{n \rightarrow \infty} a_n = a \text{ or } a_n \rightarrow a$$

if and only if for every $\varepsilon > 0$ there exists n_ε such that whenever $n \geq n_\varepsilon$,

$$|a_n - a| < \varepsilon.$$

Here a and a_n are assumed to be real numbers but the same definition holds more generally.

In words the definition says that given any measure of closeness ε , the terms of the sequence are **eventually** this close to a . Here, the word “eventually” refers to n being sufficiently large. The above definition is always the definition of what is meant by the limit of a sequence. However, in practice we usually say that something happens for n sufficiently large rather than trying to specify a particular size for how large n must be. First is a situation where the limit always exists. Nor do we determine limits by doing experiments with calculators.

Proposition 3.3.2 Let $\{a_n\}_{n=1}^{\infty}$ be an increasing sequence meaning $a_n \leq a_{n+1}$ for all n and suppose $a \equiv \sup\{a_n : n \geq 1\} < \infty$. Then $\lim_{n \rightarrow \infty} a_n = a$. A similar result holds if the sequence is decreasing and bounded below if $a \equiv \inf\{a_n : n \geq 1\}$.

Proof: For each $\varepsilon > 0$, there exists $a_n \in [a - \varepsilon, a]$ since otherwise a is not equal to what it is defined to be. Since $\{a_n\}$ is increasing, it follows that $a_n \in [a - \varepsilon, a]$ for all n large enough. Hence $\lim_{n \rightarrow \infty} a_n = a$. The situation where the sequence is decreasing and bounded below is exactly similar. ■

Next is the important theorem that the limit, if it exists, is unique.

Theorem 3.3.3 If $\lim_{n \rightarrow \infty} a_n = a$ and $\lim_{n \rightarrow \infty} a_n = \hat{a}$ then $\hat{a} = a$.

¹ Bernhard Bolzano lived from 1781 to 1848. He had an Italian father but was born in Bohemia, and he wrote in German. He was a Catholic priest and held a position in philosophy at the University of Prague. It appears that Bolzano believed in the words of Jesus and did not hesitate to enthusiastically promote what he knew was right. This got him in trouble with the political establishment of Austria. When he refused to recant, he was forced out of the university and forbidden to publish. He also displeased the Catholic hierarchy for being too rational.

Bolzano believed in absolute rigor in mathematics. He also was interested in physics, theology, and especially philosophy. His contributions in philosophy are very influential. He originated anti-psychologism also called logical objectivism which holds that logical truth exists independent of our opinions about it, contrary to the notion that truth for one person may not be truth for another. His collected writings fill some 25 volumes.

The intermediate value theorem from calculus is due to him. These days, the intermediate value theorem is considered obvious and is not discussed well in calculus texts, but Bolzano knew better and gave a proof which identified exactly what was needed instead of relying on vague intuition and geometric speculation.

Like many of the other mathematicians, he was concerned with the notion of infinitesimals which had been popularized by Leibniz. Some tried to strengthen this idea and others sought to get rid of it. They realized that something needed to be done about this fuzzy idea. Bolzano was one who contributed to removing it from calculus. He also proved the extreme value theorem in 1830's and gave the first formal $\varepsilon\delta$ description of continuity and limits. This notion of infinitesimals did not completely vanish. These days, it is called non standard analysis. It can be made mathematically respectable but not in this book.

Proof: Suppose $\hat{a} \neq a$. Then let $0 < \varepsilon < |\hat{a} - a|/2$ in the definition of the limit. It follows that there exists n_ε such that if $n \geq n_\varepsilon$, then $|a_n - a| < \varepsilon$ and $|a_n - \hat{a}| < \varepsilon$. Just let n_ε be the larger of two numbers, one which works for a and one which works for \hat{a} . Therefore, for such n ,

$$\begin{aligned} |\hat{a} - a| &\leq |\hat{a} - a_n| + |a_n - a| \\ &< \varepsilon + \varepsilon < |\hat{a} - a|/2 + |\hat{a} - a|/2 = |\hat{a} - a|, \end{aligned}$$

a contradiction. ■

Example 3.3.4 Let $a_n = \frac{1}{n^2+1}$.

Then it seems clear that $\lim_{n \rightarrow \infty} \frac{1}{n^2+1} = 0$. In fact, this is true from the definition. Let $\varepsilon > 0$ be given. Let $n_\varepsilon \geq \sqrt{\varepsilon^{-1}}$. Then if $n > n_\varepsilon \geq \sqrt{\varepsilon^{-1}}$, it follows that $n^2 + 1 > \varepsilon^{-1}$ and so $0 < \frac{1}{n^2+1} = a_n < \varepsilon$. Thus $|a_n - 0| < \varepsilon$ whenever n is this large.

Note the definition was of no use in finding a candidate for the limit. This had to be produced based on other considerations. The definition is for verifying beyond any doubt that something is the limit. It is also what must be referred to in establishing theorems which are good for finding limits.

Example 3.3.5 Let $a_n = n^2$

Then in this case $\lim_{n \rightarrow \infty} a_n$ does not exist.

Example 3.3.6 Let $a_n = (-1)^n$.

In this case, $\lim_{n \rightarrow \infty} (-1)^n$ does not exist. This follows from the definition. Let $\varepsilon = 1/2$. If there exists a limit l , then eventually, for all n large enough, $|a_n - l| < 1/2$. However, $|a_n - a_{n+1}| = 2$ and so,

$$2 = |a_n - a_{n+1}| \leq |a_n - l| + |l - a_{n+1}| < 1/2 + 1/2 = 1$$

which cannot hold. Therefore, there is no limit for this sequence.

Theorem 3.3.7 Suppose $\{a_n\}$ and $\{b_n\}$ are sequences and that

$$\lim_{n \rightarrow \infty} a_n = a \text{ and } \lim_{n \rightarrow \infty} b_n = b.$$

Also suppose x and y are in \mathbb{R} . Then

$$\lim_{n \rightarrow \infty} xa_n + yb_n = xa + yb \quad (3.1)$$

$$\lim_{n \rightarrow \infty} a_nb_n = ab \quad (3.2)$$

If $b \neq 0$,

$$\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = \frac{a}{b}. \quad (3.3)$$

Proof: The first of these claims is left for you to do. To do the second, let $\varepsilon > 0$ be given and choose n_1 such that if $n \geq n_1$ then $|a_n - a| < 1$. Then for such n , the triangle inequality implies

$$\begin{aligned} |a_n b_n - ab| &\leq |a_n b_n - a_n b| + |a_n b - ab| \\ &\leq |a_n| |b_n - b| + |b| |a_n - a| \\ &\leq (|a| + 1) |b_n - b| + |b| |a_n - a|. \end{aligned}$$

Now let n_2 be large enough that for $n \geq n_2$,

$$|b_n - b| < \frac{\varepsilon}{2(|a| + 1)}, \text{ and } |a_n - a| < \frac{\varepsilon}{2(|b| + 1)}.$$

Such a number exists because of the definition of limit. Therefore, let

$$n_\varepsilon > \max(n_1, n_2).$$

For $n \geq n_\varepsilon$,

$$\begin{aligned} |a_n b_n - ab| &\leq (|a| + 1) |b_n - b| + |b| |a_n - a| \\ &< (|a| + 1) \frac{\varepsilon}{2(|a| + 1)} + |b| \frac{\varepsilon}{2(|b| + 1)} \leq \varepsilon. \end{aligned}$$

This proves 3.2. Next consider 3.3.

Let $\varepsilon > 0$ be given and let n_1 be so large that whenever $n \geq n_1$,

$$|b_n - b| < \frac{|b|}{2}.$$

Thus for such n ,

$$\begin{aligned} \left| \frac{a_n}{b_n} - \frac{a}{b} \right| &= \left| \frac{a_n b - ab_n}{b b_n} \right| \leq \frac{2}{|b|^2} [|a_n b - ab| + |ab - ab_n|] \\ &\leq \frac{2}{|b|} |a_n - a| + \frac{2|a|}{|b|^2} |b_n - b|. \end{aligned}$$

Now choose n_2 so large that if $n \geq n_2$, then

$$|a_n - a| < \frac{\varepsilon |b|}{4}, \text{ and } |b_n - b| < \frac{\varepsilon |b|^2}{4(|a| + 1)}.$$

Letting $n_\varepsilon > \max(n_1, n_2)$, it follows that for $n \geq n_\varepsilon$,

$$\begin{aligned} \left| \frac{a_n}{b_n} - \frac{a}{b} \right| &\leq \frac{2}{|b|} |a_n - a| + \frac{2|a|}{|b|^2} |b_n - b| \\ &< \frac{2}{|b|} \frac{\varepsilon |b|}{4} + \frac{2|a|}{|b|^2} \frac{\varepsilon |b|^2}{4(|a| + 1)} < \varepsilon. \blacksquare \end{aligned}$$

Another very useful theorem for finding limits is the squeezing theorem. It is like two men supporting a drunk companion between them and the two are headed for a sink hole into which they will fall. Then the drunk companion will also fall into the hole.

Theorem 3.3.8 Suppose $\lim_{n \rightarrow \infty} a_n = a = \lim_{n \rightarrow \infty} b_n$ and $a_n \leq c_n \leq b_n$ for all n large enough. Then $\lim_{n \rightarrow \infty} c_n = a$.

Proof: Let $\varepsilon > 0$ be given and let n_1 be large enough that if $n \geq n_1$,

$$|a_n - a| < \varepsilon/2 \text{ and } |b_n - a| < \varepsilon/2.$$

Then for such n ,

$$|c_n - a| \leq |a_n - a| + |b_n - a| < \varepsilon.$$

The reason for this is that if $c_n \geq a$, then

$$|c_n - a| = c_n - a \leq b_n - a \leq |a_n - a| + |b_n - a|$$

because $b_n \geq c_n$. On the other hand, if $c_n \leq a$, then

$$|c_n - a| = a - c_n \leq a - a_n \leq |a - a_n| + |b - b_n|. \blacksquare$$

As an example, consider the following.

Example 3.3.9 Let $c_n \equiv (-1)^n \frac{1}{n}$ and let $b_n = \frac{1}{n}$, and $a_n = -\frac{1}{n}$. Then you may easily show that

$$\lim_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} b_n = 0.$$

Since $a_n \leq c_n \leq b_n$, it follows $\lim_{n \rightarrow \infty} c_n = 0$ also.

Theorem 3.3.10 $\lim_{n \rightarrow \infty} r^n = 0$. Whenever $|r| < 1$.

Proof: If $0 < r < 1$ it follows $r^{-1} > 1$. Why? Letting $\alpha = \frac{1}{r} - 1$, it follows $r = \frac{1}{1+\alpha}$. Therefore, by the binomial theorem,

$$0 < r^n = \frac{1}{(1+\alpha)^n} \leq \frac{1}{1+\alpha n}.$$

Therefore, $\lim_{n \rightarrow \infty} r^n = 0$ if $0 < r < 1$. Now in general, if $|r| < 1$, $|r^n| = |r|^n \rightarrow 0$ by the first part. \blacksquare

For sequences, it is very important to consider something called a subsequence.

Definition 3.3.11 Let $\{a_n\}$ be a sequence and let $n_1 < n_2 < n_3, \dots$ be any strictly increasing list of integers such that n_1 is at least as large as the first n for the original sequence. Then if $b_k \equiv a_{n_k}$, $\{b_k\}$ is called a subsequence of $\{a_n\}$. We usually simply write $\{a_{n_k}\}$ to denote this subsequence.

Example 3.3.12 Suppose $a_n = (n^2 + 1)$. Thus $a_1 = 2$, $a_3 = 10$, etc. If

$$n_1 = 1, n_2 = 3, n_3 = 5, \dots, n_k = 2k - 1,$$

then letting $b_k = a_{n_k}$, it follows

$$b_k = ((2k - 1)^2 + 1) = 4k^2 - 4k + 2.$$

In general, a subsequence is just as defined. You won't necessarily be able to give a formula for the k^{th} term.

Example 3.3.13 Let $a_n = n$. Then let n_k be the k^{th} prime. Then there is no formula for a_{n_k} .

An important theorem is the one which states that if a sequence converges, so does every subsequence.

Theorem 3.3.14 Let $\{x_n\}$ be a sequence with $\lim_{n \rightarrow \infty} x_n = x$ and let $\{x_{n_k}\}$ be a subsequence. Then $\lim_{k \rightarrow \infty} x_{n_k} = x$.

Proof: Let $\varepsilon > 0$ be given. Then there exists n_ε such that if $n > n_\varepsilon$, then $|x_n - x| < \varepsilon$. Suppose $k > n_\varepsilon$. Then $n_k \geq k > n_\varepsilon$ and so $|x_{n_k} - x| < \varepsilon$ showing $\lim_{k \rightarrow \infty} x_{n_k} = x$ as claimed. ■

Theorem 3.3.15 Let $\{x_n\}$ be a sequence of real numbers and suppose each $x_n \leq l$ ($\geq l$) and $\lim_{n \rightarrow \infty} x_n = x$. Then $x \leq l$ ($\geq l$). More generally, let $\{x_n\}$ and $\{y_n\}$ be two sequences such that $\lim_{n \rightarrow \infty} x_n = x$ and $\lim_{n \rightarrow \infty} y_n = y$. Then if $x_n \leq y_n$ for all n sufficiently large, then $x \leq y$.

Proof: Suppose not. Suppose that $x_n \leq l$ but $x > l$. Then for n large enough,

$$|x_n - x| < x - l$$

and so

$$x - x_n < x - l \text{ which implies } x_n > l$$

a contradiction. The case where each $x_n \geq l$ is similar. Consider now the last claim. For n large enough,

$$y - x \geq (y_n - \varepsilon) - (x_n + \varepsilon) \geq (y_n - x_n) - 2\varepsilon \geq -2\varepsilon$$

Since ε is arbitrary, it follows that $y - x \geq 0$. ■

This last step is quite typical in calculus. To show something is nonnegative, you show it is larger than every negative number.

Recall Proposition 3.3.2 about convergence of increasing and decreasing bounded sequences. These always converge in some sense.

However, many sequences are neither increasing nor decreasing. Sometimes these sequences do not have a limit. However, there are two things which always exist for any sequence of real numbers.

Suppose $\{a_n\}$ is a sequence, and let $A_n \equiv \inf \{a_k : k \geq n\}$. Then $\{A_n\}$ is an increasing sequence in the sense $A_n \leq A_{n+1}$ because the sets $\{a_k : k \geq n\}$ are getting smaller as n increases. Thus, by Proposition 3.3.2, either $\{A_n\}$ is bounded above and $\lim_{n \rightarrow \infty} A_n$ is a real number equal to $\sup_n \{A_n\}$ or they are not bounded above and in this case, we say $\liminf_{n \rightarrow \infty} A_n = \infty$. Similarly, if $B_n \equiv \sup \{a_k : k \geq n\}$, the B_n are decreasing and we can also consider $\lim_{n \rightarrow \infty} B_n$ which equals $-\infty$ if not bounded below and some real number equal to $\inf_n \{B_n\}$ otherwise. This explains the following definition.

Definition 3.3.16 Let A_n, B_n be as just described relative to a sequence $\{a_n\}$. Then

$$\begin{aligned} \liminf_{n \rightarrow \infty} a_n &\equiv \lim_{n \rightarrow \infty} A_n \equiv \lim_{n \rightarrow \infty} (\inf \{a_k : k \geq n\}) \\ \limsup_{n \rightarrow \infty} a_n &\equiv \lim_{n \rightarrow \infty} B_n \equiv \lim_{n \rightarrow \infty} (\sup \{a_k : k \geq n\}) \end{aligned}$$

When $\liminf = \limsup$, this is when the limit exists.

Lemma 3.3.17 *Let $\{a_n\}$ be a sequence in $[-\infty, \infty]$. Then $\lim_{n \rightarrow \infty} a_n$ exists if and only if*

$$\liminf_{n \rightarrow \infty} a_n = \limsup_{n \rightarrow \infty} a_n$$

and in this case, the limit equals the common value of these two numbers or $\pm\infty$ when not bounded.

Proof: Suppose first $\lim_{n \rightarrow \infty} a_n = a \in \mathbb{R}$. Then, 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 $\limsup_{n \rightarrow \infty} a_n$ and $\liminf_{n \rightarrow \infty} a_n$ are both in $[a - \varepsilon, a + \varepsilon]$, showing

$$\left| \liminf_{n \rightarrow \infty} a_n - \limsup_{n \rightarrow \infty} a_n \right| \leq 2\varepsilon.$$

Since ε is arbitrary, the two must be equal and they both must equal a . Next suppose $\lim_{n \rightarrow \infty} a_n = \infty$. By definition, this means that 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 $\liminf_{n \rightarrow \infty} a_n = \limsup_{n \rightarrow \infty} a_n = \infty$. The case for $-\infty$ is similar.

Conversely, suppose $\liminf_{n \rightarrow \infty} a_n = \limsup_{n \rightarrow \infty} a_n = a$. If $a \in \mathbb{R}$, then, from the definition, since $\limsup_{n \rightarrow \infty} a_n = a$, then for all n large enough,

$$\sup\{a_k : k \geq n\} \in (a - \varepsilon, a + \varepsilon)$$

so in particular, $a_n < a + \varepsilon$. Similarly, since $\liminf_{n \rightarrow \infty} a_n = a$, for all n large enough, $a_n > a - \varepsilon$. Thus, for all n large enough, $|a - a_n| < \varepsilon$ and so $\lim_{n \rightarrow \infty} a_n = a$. If $\liminf_{n \rightarrow \infty} a_n = \limsup_{n \rightarrow \infty} a_n = \infty$, then for $l \in \mathbb{R}$, there exists N such that for $n \geq N$, $\inf_{k \geq n} a_k > l$. Therefore, $\lim_{n \rightarrow \infty} a_n = \infty$. The case for $-\infty$ is similar. ■

Here is a useful proposition.

Proposition 3.3.18 *Let $\lim_{n \rightarrow \infty} a_n = a > 0$ and suppose that each $b_n > 0$. Then*

$$\limsup_{n \rightarrow \infty} a_n b_n = a \limsup_{n \rightarrow \infty} b_n.$$

Proof: This follows from the definition. Let $\lambda_n = \sup\{a_k b_k : k \geq n\}$. For all n large enough, $a_n \in (a - \varepsilon, a + \varepsilon)$ 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

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

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

You might think of many other similar propositions but the above will suffice.

3.4 The Nested Interval Lemma

In Russia there is a kind of doll called a matrushka doll. You pick it up and notice it comes apart in the center. Separating the two halves you find an identical doll inside. Then you notice this inside doll also comes apart in the center. Separating the two halves, you find yet another identical doll inside. This goes on quite a while until the final doll is in one piece. The nested interval lemma is like a matrushka doll except the process never stops. It involves a sequence of intervals, the first containing the second, the second containing the third, the third containing the fourth and so on. The fundamental question is whether there exists a point in all the intervals. Sometimes there is such a point and this comes from completeness.

Lemma 3.4.1 *Let $I_k = [a^k, b^k]$ and suppose that for all $k = 1, 2, \dots$, $I_k \supseteq I_{k+1}$. Then there exists a point, $c \in \mathbb{R}$ which is an element of every I_k . If the diameters (length) of these intervals, denoted as $\text{diam}(I_k)$ converges to 0, then there is a unique point in the intersection of all these intervals.*

Proof: Since $I_k \supseteq I_{k+1}$, this implies

$$a^k \leq a^{k+1}, b^k \geq b^{k+1}. \quad (3.4)$$

Consequently, if $k \leq l$,

$$a^l \leq b^l \leq b^k. \quad (3.5)$$

Now define $c \equiv \sup \{a^l : l = 1, 2, \dots\}$. By the first inequality in 3.4, and 3.5

$$a^k \leq c = \sup \{a^l : l = k, k+1, \dots\} \leq b^k \quad (3.6)$$

for each $k = 1, 2, \dots$. Thus $c \in I_k$ for every k and this proves the lemma. The reason for the last inequality in 3.6 is that from 3.5, b^k is an upper bound to $\{a^l : l = k, k+1, \dots\}$. Therefore, it is at least as large as the least upper bound.

For the last claim, suppose there are two points x, y in the intersection. Then $|x - y| = r > 0$ but eventually the diameter of I_k is less than r . Thus it cannot contain both x, y . If so, assuming $y > x$, you would have $a^k \leq x < y \leq b^k$ and so $0 < r = y - x < b^k - a^k < r$. ■

This is really quite a remarkable result and may not seem so obvious. Consider the intervals $I_k \equiv (0, 1/k)$. Then there is no point which lies in all these intervals because no negative number can be in all the intervals and $1/k$ is smaller than a given positive number whenever k is large enough. Thus the only candidate for being in all the intervals is 0 and 0 has been left out of them all. The problem here is that the endpoints of the intervals were not included, contrary to the hypotheses of the above lemma in which all the intervals included the endpoints.

3.5 Exercises

1. Find $\lim_{n \rightarrow \infty} \frac{n}{3n+4}$.
2. Find $\lim_{n \rightarrow \infty} \frac{3n^4+7n+1000}{n^4+1}$.
3. Find $\lim_{n \rightarrow \infty} \frac{2^n+7(5^n)}{4^n+2(5^n)}$.

4. Find $\lim_{n \rightarrow \infty} \sqrt{(n^2 + 6n)} - n$. **Hint:** Multiply and divide by $\sqrt{(n^2 + 6n)} + n$.
5. Find $\lim_{n \rightarrow \infty} \sum_{k=1}^n \frac{1}{10^k}$.
6. For $|r| < 1$, find $\lim_{n \rightarrow \infty} \sum_{k=0}^n r^k$. **Hint:** First show $\sum_{k=0}^n r^k = \frac{r^{n+1}}{r-1} - \frac{1}{r-1}$. Then recall Theorem 3.3.10.
7. Prove $\lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n$ exists and equals a number less than 3.
8. Prove $n^{n+1} \geq (n+1)^n$ for all integers, $n \geq 3$.
9. Find $\lim_{n \rightarrow \infty} n \sin n$ if it exists. If it does not exist, explain why it does not.
10. Recall the axiom of completeness states that a set which is bounded above has a least upper bound and a set which is bounded below has a greatest lower bound. Show that a monotone decreasing sequence which is bounded below converges to its greatest lower bound. **Hint:** Let a denote the greatest lower bound and recall that because of this, it follows that for all $\varepsilon > 0$ there exist points of $\{a_n\}$ in $[a, a + \varepsilon]$.
11. Let $A_n = \sum_{k=2}^n \frac{1}{k(k-1)}$ for $n \geq 2$. Show $\lim_{n \rightarrow \infty} A_n$ exists and find the limit. **Hint:** Show there exists an upper bound to the A_n as follows.

$$\sum_{k=2}^n \frac{1}{k(k-1)} = \sum_{k=2}^n \left(\frac{1}{k-1} - \frac{1}{k} \right) = 1 - \frac{1}{n} \leq 1.$$

12. Let $H_n = \sum_{k=1}^n \frac{1}{k^2}$ for $n \geq 2$. Show $\lim_{n \rightarrow \infty} H_n$ exists. **Hint:** Use the above problem to obtain the existence of an upper bound.
13. Let $I_n = (-1/n, 1/n)$ and let $J_n = (0, 2/n)$. The intervals, I_n and J_n are open intervals of length $2/n$. Find $\cap_{n=1}^{\infty} I_n$ and $\cap_{n=1}^{\infty} J_n$. Repeat the same problem for $I_n = (-1/n, 1/n]$ and $J_n = [0, 2/n)$.
14. Let $\{a_n\}$ be a sequence in $(-\infty, \infty)$. Let $A_k \equiv \sup \{a_n : n \geq k\}$ so that, defining $\lambda \equiv \limsup_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} A_n$, the A_n being a decreasing sequence.
 - (a) Show that in all cases, there exists $B_n < A_n$ such that B_n is increasing and $\lim_{n \rightarrow \infty} B_n = \lambda$.
 - (b) Explain why, in all cases there are infinitely many k such that $a_k \in [B_n, A_n]$. **Hint:** If for all $k \geq m > n$, $a_k \leq B_n$, then $a_k < B_m$ also and so $\sup \{a_k : k \geq m\} \leq B_m < A_m$ contrary to the definition of A_m .
 - (c) Explain why there exists a subsequence $\{a_{n_k}\}$ such that $\lim_{k \rightarrow \infty} a_{n_k} = \lambda$.
 - (d) Show that if $\gamma \in [-\infty, \infty]$ and there is a subsequence $\{a_{n_k}\}$ with $\lim_{k \rightarrow \infty} a_{n_k} = \gamma$, then $\gamma \leq \lambda$.

This shows that $\limsup_{n \rightarrow \infty} a_n$ is the largest in $[-\infty, \infty]$ such that some subsequence converges to it.

15. Formulate a similar problem which shows that for $\{a_n\}$ a sequence of real numbers, $\liminf_{n \rightarrow \infty} a_n$ is the smallest number which is obtainable as a limit of a subsequence of the original sequence.
16. Let $I_n = [n, \infty)$. Find $\cap_{n=1}^{\infty} I_n$. These intervals are not bounded.

3.6 Compactness

Do you want to understand Calculus? If you do, then the topics in this section are essential to understand. I realize these things are difficult, but they provide reasons why something exists instead of forcing us to pretend that calculus is like religion where you must accept on faith the decrees of authority figures. The presentation leads to a modern version of what was first shown early in the nineteenth century by Cauchy and Bolzano and continued later by Weierstrass as part of the effort to remove the fuzziness from calculus.

The main ideas and terminology in this section which will be used extensively are as follows.

1. A set S is closed if whenever $\lim_{n \rightarrow \infty} a_n = a$, each $a_n \in S$, then $a \in S$ also.
2. A set S is open if every point of S is an interior point of some interval contained in S .
3. A set is bounded if it is contained in some interval.
4. A set S is sequentially compact means: If $\{a_n\}$ is a sequence contained in S , there exists a subsequence which converges to a point of S . The closed and bounded sets are the same as the sequentially compact sets.

Notice how the last assertion gives the **existence** of a convergent subsequence. The existence of this subsequence is what will be needed.

3.6.1 Sequential Compactness

First I will discuss the very important concept of sequential compactness. This is a property that some sets have. A set of numbers is sequentially compact if every sequence contained in the set has a subsequence which converges to a point **in the set**. It is unbelievably useful whenever you try to understand existence theorems.

Definition 3.6.1 A set, $K \subseteq \mathbb{R}$ is sequentially compact if whenever $\{a_n\} \subseteq K$ is a sequence, there exists a subsequence, $\{a_{n_k}\}$ such that this subsequence converges to a point of K .

The following theorem is part of the Heine Borel theorem.

Theorem 3.6.2 Every closed interval $[a, b]$ is sequentially compact.

Proof: Let $\{x_n\} \subseteq [a, b] \equiv I_0$. Consider the two intervals $[a, \frac{a+b}{2}]$ and $[\frac{a+b}{2}, b]$ each of which has length $(b-a)/2$. At least one of these intervals contains x_n for infinitely many values of n . Call this interval I_1 . Now do for I_1 what was done for I_0 . Split it in half and let I_2 be the interval which contains x_n for infinitely many values of n . Continue this way obtaining a sequence of nested intervals $I_0 \supseteq I_1 \supseteq I_2 \supseteq I_3 \cdots$ where the length of I_n is $(b-a)/2^n$. Now pick n_1 such that $x_{n_1} \in I_1$, n_2 such that $n_2 > n_1$ and $x_{n_2} \in I_2$, n_3 such that $n_3 > n_2$ and $x_{n_3} \in I_3$, etc. (This can be done because in each case the intervals contain x_n for infinitely many values of n .) By the nested interval lemma there exists a point c contained in all these intervals. Furthermore, $|x_{n_k} - c| < (b-a)2^{-k}$ and so $\lim_{k \rightarrow \infty} x_{n_k} = c \in [a, b]$. ■

3.6.2 Closed and Open Sets

I have been using the terminology $[a, b]$ is a closed interval to mean it is an interval which contains the two endpoints. However, there is a more general notion of what it means to be closed. Similarly there is a general notion of what it means to be open. An example of an open set is (a, b) , an interval which is missing its end points.

Definition 3.6.3 Let U be a set of points. A point $p \in U$ is said to be an interior point if whenever $|x - p|$ is sufficiently small, it follows $x \in U$ also. The set of points, x which are closer to p than δ is denoted by

$$B(p, \delta) \equiv \{x \in \mathbb{R} : |x - p| < \delta\} = (p - \delta, p + \delta).$$

This symbol, $B(p, \delta)$ is called an open ball of radius δ or an open interval. Thus a point, p is an interior point of U if there exists $\delta > 0$ such that $p \in B(p, \delta) \subseteq U$. An **open set** is one for which every point of the set is an interior point. **Closed sets** are those which are complements of open sets. Thus H is closed means H^C is open. Here H^C is more correctly denoted as $\mathbb{R} \setminus H$, and refers to the set of all points in \mathbb{R} not in H . In general, this is what is meant by the complement of a set. S^C denotes all points not in S which are in some given universal set containing S .

Proposition 3.6.4 Every closed interval $[a, b]$ is a closed set.

Proof: The complement is $(-\infty, a) \cup (b, \infty)$ and this is clearly an open set. For example, if $x > b$, then $x \in B(x, (x - b)) \subseteq (b, \infty)$. ■

Proposition 3.6.5 If \mathcal{U} is a set whose elements are open sets, then $\cup \mathcal{U}$ is also an open set.

Proof: Suppose $x \in \cup \mathcal{U}$. Then $x \in U \in \mathcal{U}$ for some U . Then for some $\delta > 0$, the interval $(x - \delta, x + \delta) \subseteq U$ and so $(x - \delta, x + \delta) \subseteq \cup \mathcal{U}$ so $\cup \mathcal{U}$ is open. ■

Example 3.6.6 What is $[0, 1]^C$?

It consists of all points not in $[0, 1]$. Thus it is $(1, \infty) \cup (-\infty, 0)$, all points which are either to the right of 1 on the number line or to the left of 0 on the number line. Note that $[0, 1]^C$ is an open set. Thus $[0, 1]$ is a closed set.

What is an example of an open set? The simplest example is an open ball.

Proposition 3.6.7 $B(p, \delta)$ is an open set and every open interval (a, b) is an open set.

Proof: It is necessary to show that every point is an interior point. Let $x \in B(p, \delta)$. Then let $r = \delta - |x - p|$. It follows $r > 0$ because it is given that $|x - p| < \delta$. Now consider $z \in B(x, r)$.

$$|z - p| \leq |z - x| + |x - p| < r + |x - p| = \delta - |x - p| + |x - p| = \delta$$

and so $z \in B(p, \delta)$. That is $B(x, r) \subseteq B(p, \delta)$. Since x was arbitrary, this has shown that every point of the ball is an interior point. Thus the ball is an open set. Now $(a, b) = B(\frac{a+b}{2}, \frac{b-a}{2})$. ■

Definition 3.6.8 Let A be any nonempty set and let x be a point. Then x is said to be a limit point of A if for every $r > 0$, $B(x, r)$ contains a point of A which is not equal to x .

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. I think the version in the proposition is a little easier to use.

Proposition 3.6.9 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 $\{a_1, \dots, a_n\}$ have been chosen none equal to x and with no repeats in the list, let $0 < r_n < \min\left(\frac{1}{n}, \min\{|a_i - x|, i = 1, 2, \dots, n\}\right)$. 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 . ■

Example 3.6.10 Consider $A = \mathbb{N}$, the positive integers. Then none of the points of A is a limit point of A because if $n \in A$, $B(n, 1/10)$ contains no points of \mathbb{N} which are not equal to n .

Example 3.6.11 Consider $A = (a, b)$, an open interval. This is an open set. Indeed, it equals the open ball $B\left(\frac{a+b}{2}, \frac{b-a}{2}\right)$ centered at the midpoint of the interval $\frac{a+b}{2}$, having radius half the length of the interval.

Theorem 3.6.12 The following are equivalent.

1. A is closed
2. If $\{a_n\}_{n=1}^{\infty}$ is a sequence of points of A and $\lim_{n \rightarrow \infty} a_n = a$, then $a \in A$.
3. A contains all of its limit points.

If a is a limit point, then there is a sequence of distinct points of A none of which equal a which converges to a .

Proof: 1. \Leftrightarrow 2. Say A is closed and $a_n \rightarrow a$ where each $a_n \in A$. If $a \notin A$, then there exists $\varepsilon > 0$ such that $B(a, \varepsilon) \cap A = \emptyset$. But then a_n fails to converge to a so $a \in A$. Conversely, if 2. holds and $x \notin A$, $B\left(x, \frac{1}{n}\right)$ must fail to contain any points of A for some $n \in \mathbb{N}$ because if not, you could pick $a_n \in B\left(x, \frac{1}{n}\right) \cap A$ and obtain $\lim_{n \rightarrow \infty} a_n = x$ which would give $x \in A$ by 2. Thus A^C is open and A is closed.

2. \Rightarrow 3. Say a is a limit point of A . Then for each $n \in \mathbb{N}$, $B\left(a, \frac{1}{n}\right)$ contains infinitely many points of A . Pick $a_1 \in A \cap B(a, 1)$, $a_1 \neq a$. If a_1, \dots, a_{n-1} have been chosen, $a_k \in B\left(a, \frac{1}{k}\right)$ no $a_k = a$, let $a_n \in B\left(a, \frac{1}{n}\right) \cap A$ and a_n is none of a_1, \dots, a_{n-1}, a . Then $\lim_{n \rightarrow \infty} a_n = a$ and so $a \in A$ by 2. Also, this sequence consists of distinct points, none of which equal a . This shows the last claim.

3. \Rightarrow 1. We need to show that A^C is open. Let $x \in A^C$. By 3. x cannot be a limit point. Hence there exists $B(x, r)$ which contains at most finitely many points of A . Since $x \in A^C$, none of these are equal to x . Hence, making r still smaller, one can avoid all of these points. Thus the modified r has the property that $B(x, r)$ contains no points of A and so A is closed. ■

Note that part of this theorem says that a set A having all its limit points is the same as saying that whenever a sequence of points of A converges to a point a , then it follows $a \in A$. In other words, closed is the same as being closed with respect to containing all limits of sequences of points of A .

Corollary 3.6.13 *Let A be a nonempty set and denote by A' the set of limit points of A . Then $A \cup A'$ is a closed set and it is the smallest closed set containing A .*

Proof: Is it the case that $(A \cup A')^C$ is open? This is what needs to be shown if the given set is closed. Let $p \notin A \cup A'$. Then since p is neither in A nor a limit point of A , there exists $B(p, r)$ such that $B(p, r) \cap A = \emptyset$. Therefore, $B(p, r) \cap A' = \emptyset$ also. This is because if $z \in B(p, r) \cap A'$, then $B(z, r - |p - z|) \subseteq B(p, r)$ and this smaller ball contains points of A since z is a limit point. This contradiction shows that $B(p, r) \cap A' = \emptyset$ as claimed. Hence $(A \cup A')^C$ is open because p was an arbitrary point of $(A \cup A')^C$ and $A \cup A'$ is closed as claimed.

Now suppose $C \supseteq A$ and C is closed. Then if p is a limit point of A , it follows from Theorem 3.6.12 that there exists a sequence of distinct points of A converging to p . Since C is closed, and these points of A are all in C , it follows that $p \in C$. Hence $C \supseteq A \cup A'$. ■

Theorem 3.6.14 *If K is sequentially compact and if H is a closed subset of K then H is sequentially compact.*

Proof: Let $\{x_n\} \subseteq H$. Then since K is sequentially compact, there is a subsequence, $\{x_{n_k}\}$ which converges to a point, $x \in K$. But these x_{n_k} are in the closed set H and so $x \in H$ also thanks to Theorem 3.6.12. ■

Thus every closed subset of a closed interval is sequentially compact. This is equivalent to the following corollary in which a set is said to be bounded if it is contained in some closed interval of finite length.

Corollary 3.6.15 *Every closed and bounded set in \mathbb{R} is sequentially compact.*

Proof: Let H be a closed and bounded set in \mathbb{R} . Then $H \subseteq [a, b]$ for some interval of the form $[a, b]$. Therefore, H is sequentially compact. ■

In fact, one can go the other way. First is a simple lemma.

Lemma 3.6.16 *If $\lim_{n \rightarrow \infty} x_n = x$, then the sequence $\{x_n\}$ is contained in some interval so it is bounded.*

Proof: By definition, there exists N such that if $n \geq N$, then $|x - x_n| < 1$. By triangle inequality, $|x_n| \leq (|x| - 1, |x| + 1)$, $n \geq N$. Let $M \equiv \max\{|x_k| : k \leq N\}$. Then for all n , $|x_n| \in ((|x| - (1 + M)), |x| + (1 + M))$. ■

Proposition 3.6.17 *A nonempty set $K \subseteq \mathbb{R}$ is sequentially compact if and only if it is closed and bounded.*

Proof: From the above corollary, if the set is closed and bounded, then it is sequentially compact. Suppose now that K is sequentially compact. Why is it closed and bounded? If it is not bounded, then you could pick $\{k_n\}_{n=1}^{\infty}$ such that $|k_n| \geq n$. Since K is sequentially compact, it follows that there is a subsequence, $\{k_{n_j}\}$ which satisfies $\lim_{j \rightarrow \infty} k_{n_j} = k \in K$. But then this sub sequence would be contained in some interval which is impossible from

the construction. Thus K is bounded. Why must K be closed? Suppose K fails to contain p where $p = \lim_{n \rightarrow \infty} p_n, p_n \in K$. A subsequence $\{p_{n_k}\}$ must converge to a point of K . But this subsequence must converge to p by Theorem 3.3.14 which is a contradiction. By Theorem 3.6.12 K is closed. ■

3.7 Cauchy Sequences

Definition 3.7.1 A sequence $\{x_k\}_{k=1}^{\infty}$ is called a *Cauchy sequence* if for any $\varepsilon > 0$ there exists n_ε such that whenever $m, n \geq n_\varepsilon$, it follows that $|x_n - x_m| < \varepsilon$. In other words, the terms of the sequence “bunch up”.

I will be vague about the context of the following fundamental proposition because it applies in far greater generality than \mathbb{R} . You can think of the sequence being in \mathbb{R} because this is the main example of interest here. Part 1. is especially useful in more general contexts.

Proposition 3.7.2 If $\{x_n\}$ is a Cauchy sequence, then

1. If a subsequence $\{x_{n_k}\}_{k=1}^{\infty}$ converges to x , it follows that $\lim_{n \rightarrow \infty} x_n = x$.
2. If $\lim_{n \rightarrow \infty} x_n = x$, then $\{x_n\}$ must be a Cauchy sequence.
3. Every Cauchy sequence is bounded.

Proof: Consider 1. There exists n_ε such that if $n, m > n_\varepsilon$, then $|x_n - x_m| < \varepsilon/3$. There also exists k_ε such that if $k > k_\varepsilon$, then $|x - x_{n_k}| < \varepsilon/3$. Now let $k > \max(k_\varepsilon, n_\varepsilon)$. Then $n_k \geq k > \max(k_\varepsilon, n_\varepsilon)$ and so $|x - x_k| \leq |x - x_{n_k}| + |x_{n_k} - x_k| < \frac{\varepsilon}{3} + \frac{\varepsilon}{3} < \varepsilon$ so $\lim_{n \rightarrow \infty} x_n = x$.

2. As to the next claim, there is N such that if $m \geq N$, then $|x - x_m| < \varepsilon/2$. If $m, n > N$, then $|x_m - x_n| \leq |x_m - x| + |x - x_n| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$ and so any convergent sequence is a Cauchy sequence.

3. Finally, if $\{x_n\}$ is a Cauchy sequence, then there exists N such that if $m, n \geq N$, then $|x_m - x_n| \leq 1$. In particular, $|x_n - x_N| \leq 1$ and so $|x_n| \leq |x_N| + 1$. Now for any k , $|x_k| \leq \max(|x_N| + 1, |x_i|, i = 1, 2, \dots, N)$. ■

Theorem 3.7.3 Let $\{x_n\}$ be a Cauchy sequence in \mathbb{R} . Then it converges. Conversely, if a sequence $\{x_n\}$ converges, then the sequence is a Cauchy sequence.

Proof: Since $\{x_n\}$ is a Cauchy sequence, it is bounded by Proposition 3.7.2 so is contained in some closed interval $[-a, a]$, a sequentially compact set from the above Theorem 3.6.2. Therefore, there is a subsequence $\{x_{n_k}\}$ such that $\lim_{k \rightarrow \infty} x_{n_k} = x \in [-a, a]$. By Proposition 3.7.2, the original Cauchy sequence converges to x . The second claim is from Proposition 3.7.2. ■

3.8 Exercises

1. Show the intersection of any collection of closed sets is closed and the union of any collection of open sets is open.
2. Show that if H is closed and U is open, then $H \setminus U$ is closed.

3. Show the finite intersection of any collection of open sets is open. Next show that $U \setminus H$ is open if U is open and H is closed.
4. Show the finite union of any collection of closed sets is closed.
5. Suppose $\{H_n\}_{n=1}^N$ is a finite collection of sets and suppose x is a limit point of $\bigcup_{n=1}^N H_n$. Show x must be a limit point of at least one H_n .
6. Give an example of a set of closed sets whose union is not closed.
7. Give an example of a set of open sets whose intersection is not open.
8. Give an example of a set of open sets whose intersection is a closed interval.
9. Give an example of a set of closed sets whose union is open.
10. Give an example of a set of closed sets whose union is an open interval.
11. Give an example of a set of open sets whose intersection is closed.
12. Give an example of a set of open sets whose intersection is the natural numbers.
13. Explain why \mathbb{R} and \emptyset are sets which are both open and closed when considered as subsets of \mathbb{R} .
14. Let U be any open set in \mathbb{R} . Show that every point of U is a limit point of U .
15. Suppose $\{K_n\}$ is a sequence of sequentially compact nonempty sets which have the property that $K_n \supseteq K_{n+1}$ for all n . Show there exists a point in the intersection of all these sets, denoted by $\bigcap_{n=1}^{\infty} K_n$. This is like the nested interval lemma.
16. Now suppose $\{K_n\}$ is a sequence of sequentially compact nonempty sets which have the finite intersection property, every finite subset of $\{K_n\}$ has nonempty intersection. Show there exists a point in $\bigcap_{n=1}^{\infty} K_n$.
17. Show that any finite union of sequentially compact sets is sequentially compact.
18. Completeness of \mathbb{R} was expressed earlier in terms of the existence of a least upper bound and greatest lower bound for any bounded set. This was the version of completeness used by Bolzano. From this, it was shown that a closed and bounded set is sequentially compact, Proposition 3.6.17. Show first that every bounded sequence in \mathbb{R} has a convergent subsequence. This is called the Weierstrass Bolzano theorem. Prove from this that every Cauchy sequence in \mathbb{R} must converge. **Hint:** For the first part, the bounded sequence is contained in some closed interval $[a, b]$ which is sequentially compact. For the second part, show that any Cauchy sequence is bounded so it has a convergent subsequence. Then use Proposition 3.7.2.
19. \uparrow From the above problem, the Bolzano version of completeness implies every Cauchy sequence converges. Now suppose you know that every Cauchy sequence converges. First use this to prove the nested interval lemma in the case that the diameters of the intervals converge to 0. Next show the existence of a least upper bound to a nonempty set which is bounded above. Thus convergence of any Cauchy sequence implies the version of completeness involving existence of least upper bounds. Thus convergence of Cauchy sequences is equivalent to the standard definition. **Hint:** For the

first part, pick a point $p_k \in I_k$ where the I_k are the nested closed intervals having diameter converging to 0. This is a Cauchy sequence and each interval is a closed set. Thus the limit of $\{p_k\}$ must be in I_k for each k . For the second part, if A is a nonempty set bounded above, show there exists a sequence of intervals whose diameter converges to 0 which have the left end in A and the right end an upper bound for A . Then apply the nested interval lemma.

20. Suppose you have a sequentially compact set K and suppose that \mathcal{C} is a set whose elements are open sets such that every point of K is contained in some set of \mathcal{C} . Show the existence of a number $\delta > 0$ which is a positive number such that for every $x \in K$, $B(x, \delta)$ is contained in some set of \mathcal{C} . This is called a Lebesgue number. **Hint:** If there is no Lebesgue number, then for each $n \in \mathbb{N}$, $1/n$ is not a Lebesgue number. Hence there exists $x_n \in K$ such that $B(x_n, 1/n)$ is not contained in any single set of \mathcal{C} . Extract a convergent subsequence, still denoted as $x_n \rightarrow x$. Then $B(x, \delta)$ is contained in a single set of \mathcal{C} . Isn't it the case that $B(x_n, 1/n)$ is contained in $B(x, \delta)$ for all n large enough? Isn't this a contradiction?
21. \uparrow A set \mathcal{C} whose entries are open sets is called an open cover of K if every point of K is contained in some set of \mathcal{C} . This is written as $K \subseteq \cup \mathcal{C}$. Recall the meaning of $\cup \mathcal{C}$. It is the set of all elements of some set of \mathcal{C} . The real definition of compactness is as follows: A nonempty set K is compact if and only if whenever \mathcal{C} is an open cover of K , there are finitely many sets in \mathcal{C} whose union contains K . Show that any sequentially compact set is compact. **Hint:** Get δ a Lebesgue number and show that there are finitely many points $x_i \in K$ such that $K \subseteq \cup_{i=1}^n B(x_i, \delta)$ since otherwise, you could obtain a sequence which has no converging subsequence.
22. Next show that if K is a nonempty compact set, then it must also be sequentially compact. **Hint:** If not, there would be a sequence of points of K with no subsequence converging to a point of K . Explain why this sequence can't have a limit point in K and cannot repeat infinitely often. Then show $H \cup \cup_{k=n}^{\infty} \{x_k\}$ is a closed set where here H is the set of all limit points of the sequence. Then let $U_n \equiv (H \cup \cup_{k=n}^{\infty} \{x_k\})^C$. No finite collection of the U_n can cover K .
23. Suppose \mathcal{K} is a set whose entries are nonempty compact sets. Also suppose there is a nonempty intersection of any finite collection of sets of \mathcal{K} . This is called the finite intersection property. Verify that there is a point which is contained in every set of \mathcal{K} . That is $\cap \mathcal{K} \neq \emptyset$. This is an amazing result. It actually follows right away from the definition of compactness. Recall the meaning of $\cap \mathcal{K}$ as the set of all elements which are in every set of \mathcal{K} .
24. Show that the set of limit points of a nonempty set is a closed set.
25. Let $[a, b]$ be an interval and suppose $a = x_0 < x_1 < \dots < x_n = b$. Then this ordered list of intermediate points (x_0, x_1, \dots, x_n) is called a partition of the interval $[a, b]$. Letting $f : [a, b] \rightarrow \mathbb{R}$ be a bounded function, let $M_i \equiv \sup \{f(x) : x \in [x_{i-1}, x_i]\}$ and $m_i \equiv \inf \{f(x) : x \in [x_{i-1}, x_i]\}$. Then $U(f, P)$ defined as $\sum_{i=1}^n M_i (x_i - x_{i-1})$ is called an upper sum and $L(f, P)$ defined as $\sum_{i=1}^n m_i (x_i - x_{i-1})$ is called a lower sum. Show that if P, Q are two partitions and if $P \subseteq Q$, then $U(f, P) \geq U(f, Q)$ and $L(f, P) \leq L(f, Q)$. **Hint:** To do this, show that the inequalities result from adding in one point to P to get Q .

26. ↑Now show that for P, Q any two partitions,

$$U(f, P) \geq U(f, P \cup Q) \geq L(f, P \cup Q) \geq L(f, Q).$$

Next use the above problem to verify that for

$$\bar{I} \equiv \inf \{U(f, P) : P \text{ is a partition}\}$$

and $\underline{I} \equiv \sup \{L(f, P) : P \text{ is a partition}\}$, it follows that $\underline{I} \leq \bar{I}$. When these two are equal, we say that the function is integrable and we write $\int_a^b f(x) dx$ for the common value or more simply $\int_a^b f dx$.

27. ↑Show that any decreasing function defined on $[a, b]$ is integrable. Decreasing means that if $x > \hat{x}$, then $f(\hat{x}) \geq f(x)$. The function is increasing if $f(x) \geq f(\hat{x})$. Next show that any increasing function defined on $[a, b]$ is integrable.
28. ↑Suppose $[a, b]$ is an interval and f is a bounded real valued function defined on this interval and that there is a partition $a = z_0 < z_1 < \cdots < z_n = b$ such that f is either increasing or decreasing on each sub interval $[z_{i-1}, z_i]$. Show that then $\int_a^b f dx$ exists. Thus all reasonable bounded functions are integrable.
29. Suppose a bounded real valued function f is integrable on $[a, c]$ and that $a < b < c$. Show that the restrictions of this function to $[a, b]$ and $[b, c]$ are integrable on these intervals and in fact,

$$\int_a^b f dx + \int_b^c f dx = \int_a^c f dx$$

Also explain why the function is integrable on any interval which is a subset of $[a, c]$.

30. Define $\int_b^a f dx \equiv -\int_a^b f dx$. Suppose f is integrable on

$$[\min(p, q, r), \max(p, q, r)].$$

Then show $\int_p^q f dx + \int_q^r f dx = \int_p^r f dx$.

3.9 Videos

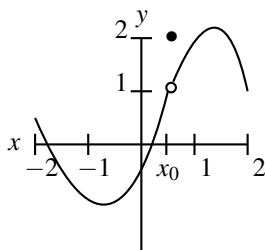
[1 sequences and functions](#) [2 limits of sequences](#)
[3 open, closed and compact](#) [Darboux integral](#)

Chapter 4

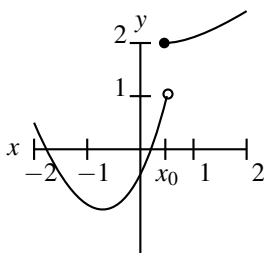
Continuous Functions and Limits of Functions

Earlier the idea of a function was described and the principal ideas related to functions of natural numbers (sequences) were presented. However, the concept of function is far too general to be useful in calculus. There are various ways to restrict the concept in order to study something interesting and the types of restrictions considered depend very much on what you find interesting. In calculus, the most fundamental restriction made is to assume the functions are continuous. Continuous functions are those in which a sufficiently small change in x results in a small change in $f(x)$. They rule out things which could never happen physically. For example, it is not possible for a car to jump from one point to another instantly. Making this restriction precise turns out to be surprisingly difficult and if you want to understand calculus, you must seek to master these difficult ideas. **There are no short cuts which will suffice.**

Before giving the careful mathematical definitions, here are examples of graphs of functions which are not continuous at the point x_0 .



You see, there is a hole in the picture of the graph of this function and instead of filling in the hole with the appropriate value, $f(x_0)$ is too large. This is called a removable discontinuity because the problem can be fixed by redefining the function at the point x_0 . Here is another example.



You see from this picture that there is no way to get rid of the jump in the graph of this function by simply redefining the value of the function at x_0 . That is why it is called a nonremovable discontinuity or jump discontinuity. Now that pictures have been given of what it is desired to eliminate, it is time to give the precise definition.

The definition which follows, due to Cauchy,¹ Bolzano, and Weierstrass² is the precise way to exclude the sort of behavior described above and all statements about continuous functions must ultimately rest on this definition from now on.

Definition 4.0.1 A function $f : D(f) \subseteq \mathbb{R} \rightarrow \mathbb{R}$ is continuous at $x \in D(f)$ if for each $\varepsilon > 0$ there exists $\delta > 0$ such that whenever $y \in D(f)$ and $|y - x| < \delta$ it follows that

$$|f(x) - f(y)| < \varepsilon.$$

A function f is continuous if it is continuous at every point of $D(f)$.

In sloppy English this definition says roughly the following: A function f is continuous at x when it is possible to make $f(y)$ as close as desired to $f(x)$ provided y is taken close enough to x . In fact this statement in words is pretty much the way Bolzano described it. Cauchy described it similarly, if his description is interpreted appropriately. The completely rigorous definition above is associated more with Weierstrass. This definition does indeed

¹Augustin Louis Cauchy 1789-1857 was the son of a lawyer who was married to an aristocrat. He was born in France just after the fall of the Bastille and his family fled the reign of terror and hid in the countryside till it was over. Cauchy was educated at first by his father who taught him Greek and Latin. Eventually Cauchy learned many languages.

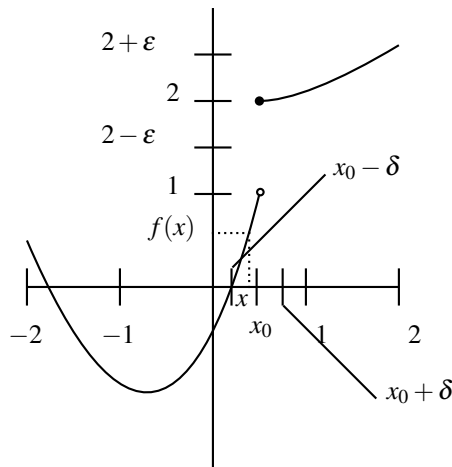
After the reign of terror, the family returned to Paris and Cauchy studied at the university to be an engineer but became a mathematician although he made fundamental contributions to physics and engineering. Cauchy was one of the most prolific mathematicians who ever lived. He wrote several hundred papers which fill 24 volumes. He also did research on many topics in mechanics and physics including elasticity, optics and astronomy. More than anyone else, Cauchy invented the subject of complex analysis. He is also credited with giving the first rigorous use of continuity in terms of ε, δ arguments in some of his work, although he clung to the notion of infinitesimals. He might have his name associated with more important topics in mathematics and engineering than any other person. He was a devout Catholic, a royalist, adhering to the Bourbons, and a man of integrity and principle, according to his understanding.

He married in 1818 and lived for 12 years with his wife and two daughters in Paris till the revolution of 1830. Cauchy was a "Legitimist" and refused to take the oath of allegiance to the new ruler, Louis Philippe because Louis was not sufficiently Bourbon, and ended up leaving his family and going into exile for 8 years. It wasn't the last time that he refused to take such an oath.

Notwithstanding his great achievements he was not a popular teacher.

²Wilhelm Theodor Weierstrass 1815-1897 brought calculus to essentially the state it is in now. When he was a secondary school teacher, he wrote a paper which was so profound that he was granted a doctor's degree. He made fundamental contributions to partial differential equations, complex analysis, calculus of variations, number theory, and many other topics. He also discovered some pathological examples such as nowhere differentiable continuous functions. Cauchy and Bolzano both had used the main ideas of the $\varepsilon - \delta$ definition, but it became well established by Weierstrass who is usually given credit for this definition. Rigorous calculus as we have it now developed over a long period of time.

rule out the sorts of graphs drawn above. Consider the second non-removable discontinuity. The removable discontinuity case is similar.



For the ε shown, you can see from the picture that no matter how small you take δ , there will be points x , between $x_0 - \delta$ and x_0 where $f(x) < 2 - \varepsilon$. In particular, for these values of x , $|f(x) - f(x_0)| > \varepsilon$. Therefore, the definition of continuity given above excludes the situation in which there is a jump in the function. Similar reasoning shows it excludes the removable discontinuity case as well. There are many ways a function can fail to be continuous and it is impossible to list them all by drawing pictures. This is why it is so important to use the definition or something equivalent to it. Here is a useful re-formulation in terms of sequences. This re-formulation seems to be easier for most of us to use. I think it is because it makes fewer explicit references to quantifiers and is symbolically easier to write although the quantifiers are already hidden in the statements about sequences.

Theorem 4.0.2 *A function $f : D(f) \subseteq \mathbb{R} \rightarrow \mathbb{R}$ is continuous at $x \in D(f)$ if and only if whenever $x_n \rightarrow x$ with $x_n \in D(f)$, it follows that $f(x_n) \rightarrow f(x)$.*

Proof: \Rightarrow Suppose f is continuous at $x \in D(f)$ and $x_n \rightarrow x$. If $\varepsilon > 0$ is given, then by assumption of continuity, there exists $\delta > 0$ such that if $|y - x| < \delta$, then $|f(y) - f(x)| < \varepsilon$. However, since $x_n \rightarrow x$, it follows that for all n large enough, $|x_n - x| < \delta$ and so $|f(x_n) - f(x)| < \varepsilon$. Thus $f(x_n) \rightarrow f(x)$.

\Leftarrow Suppose the sequence condition holds at x . Why is f continuous at x ? If this were not the case, then there would exist $\varepsilon > 0$ and x_n with $|x_n - x| < 1/n$ but $|f(x_n) - f(x)| \geq \varepsilon$. However, $x_n \rightarrow x$ and so $f(x_n) \rightarrow f(x)$ so for large enough n , $|f(x) - f(x_n)| < \varepsilon$, a contradiction. ■

Because of this theorem, I will use either of the equivalent definitions without comment in what follows.

The other thing to notice is that the concept of continuity as described in the definition is a point property. That is to say it is a property which a function may or may not have at a single point. Here is an example.

Example 4.0.3 Let $f(x) = \begin{cases} x & \text{if } x \text{ is rational} \\ 0 & \text{if } x \text{ is irrational} \end{cases}$. This function is continuous at $x = 0$ and nowhere else.

Let $x_n \rightarrow 0$. Then $|f(x_n)| \leq |x_n|$ and so $f(x_n) \rightarrow 0$. Thus it is continuous at 0. If $x \neq 0$, then if x is irrational, you could pick a sequence of rational numbers $x_n \rightarrow x$. Then $f(x_n) \rightarrow x \neq 0$ but $f(0) = 0$. If x is rational, then let $x_n \rightarrow x$ where x_n is irrational. Then $f(x_n) = 0 \rightarrow 0$ but $f(x) = x \neq 0$ so this is not continuous at any other point than 0.

Here is another example.

Example 4.0.4 Show the function $f(x) = 3x + 10$ is continuous at $x = -3$.

Let $x_n \rightarrow -3$. Then by the limit theorems for sequences, $3x_n + 10 \rightarrow 3(-3) + 10 = f(-3)$ so this is continuous.

Here is another example.

Example 4.0.5 Show the function $f(x) = \sqrt{x}$ is continuous at $x = 5$.

Note $f(5) = \sqrt{5}$ and so $|f(x) - f(5)| = |\sqrt{x} - \sqrt{5}|$. For x positive, $|\sqrt{x} - \sqrt{5}| \leq \frac{|x-5|}{\sqrt{x}+\sqrt{5}} \leq \frac{|x-5|}{\sqrt{5}}$. Now let $x_n \rightarrow 5$. Eventually x_n is positive and so $|\sqrt{x_n} - \sqrt{5}| \leq \frac{|x_n-5|}{\sqrt{5}}$ and the expression on the right converges to 0 as $n \rightarrow \infty$.

The following is a useful theorem which can remove the need to constantly use the ϵ, δ definition given above.

Theorem 4.0.6 The following assertions are valid

1. The function $af + bg$ is continuous at x when f, g are continuous at $x \in D(f) \cap D(g)$ and $a, b \in \mathbb{R}$.
2. If f and g are each continuous at x in the domains of both f and g , then fg is continuous at x . If, in addition to this, $g(x) \neq 0$, then f/g is continuous at x .
3. If f is continuous at x , $f(x) \in D(g) \subseteq \mathbb{R}$, and g is continuous at $f(x)$, then $g \circ f$ is continuous at x .
4. The function $f : \mathbb{R} \rightarrow \mathbb{R}$, given by $f(x) = |x|$ is continuous.

Proof: 1. Let $x_n \rightarrow x, x_n \in D(f) \cap D(g)$ so the combination of functions makes sense. Then $f(x_n) \rightarrow f(x), g(x_n) \rightarrow g(x)$ and so from the limit theorems for sequences, Theorem 3.3.7, $af(x_n) + bg(x_n) \equiv (af + bg)(x_n) \rightarrow af(x) + bg(x) \equiv (af + bg)(x)$.

2. Letting $x_n \rightarrow x$, continuity of f, g at x implies $f(x_n) \rightarrow f(x), g(x_n) \rightarrow g(x)$ and so from the limit theorems for sequences, $fg(x_n) \equiv f(x_n)g(x_n) \rightarrow f(x)g(x) \equiv fg(x)$.

3. If $x_n \rightarrow x$, then $f(x_n) \rightarrow f(x)$ and so $g \circ f(x_n) \equiv g(f(x_n)) \rightarrow g(f(x)) \equiv g \circ f(x)$.

4. From the triangle inequality, $||x| - |x_n|| \leq |x_n - x|$ so if $x_n \rightarrow x$, then $|x_n| \rightarrow |x|$ and so $|\cdot|$ is continuous. ■

Theorem 4.0.7 Suppose $f : D(f) \rightarrow \mathbb{R}$ is continuous at $x \in D(f)$ and suppose $f(x_n) \leq l$ ($\geq l$) where $\{x_n\}$ is a sequence of points of $D(f)$ which converges to x . Then $f(x) \leq l$ ($\geq l$).

Proof: Since $f(x_n) \leq l$ and f is continuous at x , it follows from Theorem 3.3.15 and Theorem 4.0.2, $f(x) = \lim_{n \rightarrow \infty} f(x_n) \leq l$. The other case is entirely similar. ■

The following theorem is a summary of what was shown above. I am being purposely vague about the domain of the function and its range because this theorem is a general result which holds whenever it makes sense.

Theorem 4.0.8 *Let f be a function defined on $D(f)$. The following are equivalent.*

1. f is continuous on $D(f)$
2. For every $\varepsilon > 0$ and $x \in D(f)$ there exists $\delta > 0$ such that if $|y - x| < \delta$ and $y \in D(f)$, then $|f(x) - f(y)| < \varepsilon$.
3. For every $x \in D(f)$, if $x_n \rightarrow x$ where each $x_n \in D(f)$, then $f(x) = \lim_{n \rightarrow \infty} f(x_n)$.

Proof: The first two conditions are equivalent by definition. The last two are equivalent by Theorem 4.0.2. ■

The next theorem gives an equivalence between f being continuous on $D(f)$ and something involving open intervals. It is a special case of something more general but I am only giving what will be needed later.

4.1 An Equivalent Formulation of Continuity

This is about a formulation of continuity in terms of inverse images of open intervals.

Theorem 4.1.1 *Let $f : (a, b) \rightarrow \mathbb{R}$. Then f is continuous at every point of (a, b) if and only if $f^{-1}(c, d)$ is an open subset of (a, b) for any $c < d$.*

Proof: \Rightarrow Let $x \in f^{-1}(c, d) \cap (a, b)$. If there is no open interval containing x which is contained in $f^{-1}(c, d) \cap (a, b)$, then letting n be large enough that $I_n \equiv (x - \frac{1}{n}, x + \frac{1}{n}) \subseteq (a, b)$, it must be the case that I_n has a point x_n which is not in $f^{-1}(c, d)$ meaning that either $f(x_n) \geq d$ or $f(x_n) \leq c$. Thus there is a subsequence $\{x_{n_k}\}$ for which $f(x_{n_k}) \leq c$ or for which $f(x_{n_k}) \geq d$. Suppose the latter case. The other is similar. Then from Theorem 4.0.7, $f(x) = \lim_{k \rightarrow \infty} f(x_{n_k}) \geq d$ contrary to $c < f(x) < d$ which holds because $x \in f^{-1}(c, d)$. Thus $f^{-1}(c, d) \cap (a, b)$ must be open after all. Of course, if there is no $x \in f^{-1}(c, d) \cap (a, b)$, then $f^{-1}(c, d) \cap (a, b) = \emptyset$ which is open.

\Leftarrow Let $x \in (a, b)$ and suppose $f^{-1}(c, d) \cap (a, b)$ is always open. Why is f continuous at x ? If not, there exists $x_n \rightarrow x$ but $f(x_n) \not\rightarrow f(x)$, the symbol $\not\rightarrow$ meaning it doesn't converge. It follows there exists $\varepsilon > 0$ and a subsequence $\{x_{n_k}\}$ such that $f(x_{n_k}) \notin (f(x) - \varepsilon, f(x) + \varepsilon) \equiv I_\varepsilon$. But $x \in f^{-1}(I_\varepsilon) \cap (a, b)$ and this is open so eventually $x_{n_k} \in f^{-1}(I_\varepsilon) \cap (a, b)$ and so $f(x_{n_k}) \in I_\varepsilon$ after all. Thus $x_n \rightarrow x \Rightarrow f(x_n) \rightarrow f(x)$ and so f is continuous at x after all. ■

4.2 Exercises

1. Let $f(x) = 2x + 7$. Show f is continuous at every point x . **Hint:** You need to let $\varepsilon > 0$ be given. In this case, you should try $\delta \leq \varepsilon/2$. Note that if one δ works in the definition, then so does any smaller δ .
2. Suppose $D(f) = [0, 1] \cup \{9\}$ and $f(x) = x$ on $[0, 1]$ while $f(9) = 5$. Is f continuous at the point, 9? Use whichever definition of continuity you like.
3. Let $f(x) = x^2 + 1$. Show f is continuous at $x = 3$. **Hint:**

$$|f(x) - f(3)| = |x^2 + 1 - (9 + 1)| = |x + 3||x - 3|.$$

Thus if $|x-3| < 1$, it follows from the triangle inequality, $|x| < 1+3=4$ and so $|f(x)-f(3)| < 4|x-3|$. Now complete the argument by letting $\delta \leq \min(1, \varepsilon/4)$. The symbol, \min means to take the minimum of the two numbers in the parenthesis.

4. Let $f(x) = 2x^2 + 1$. Show f is continuous at $x = 1$.
5. Let $f(x) = x^2 + 2x$. Show f is continuous at $x = 2$. Then show it is continuous at every point.
6. Let $f(x) = |2x+3|$. Show f is continuous at every point. **Hint:** Review the two versions of the triangle inequality for absolute values.
7. Let $f(x) = \frac{1}{x^2+1}$. Show f is continuous at every value of x .

8. If $x \in \mathbb{R}$, show there exists a sequence of rational numbers, $\{x_n\}$ such that $x_n \rightarrow x$ and a sequence of irrational numbers, $\{x'_n\}$ such that $x'_n \rightarrow x$. Now consider the following function.

$$f(x) = \begin{cases} 1 & \text{if } x \text{ is rational} \\ 0 & \text{if } x \text{ is irrational} \end{cases}.$$

Show using the sequential version of continuity in Theorem 4.0.2 that f is discontinuous at every point.

9. If $x \in \mathbb{R}$, show there exists a sequence of rational numbers, $\{x_n\}$ such that $x_n \rightarrow x$ and a sequence of irrational numbers, $\{x'_n\}$ such that $x'_n \rightarrow x$. Now consider the following function.

$$f(x) = \begin{cases} x & \text{if } x \text{ is rational} \\ 0 & \text{if } x \text{ is irrational} \end{cases}.$$

Show using the sequential version of continuity in Theorem 4.0.2 that f is continuous at 0 and nowhere else.

10. Suppose y is irrational and $y_n \rightarrow y$ where y_n is rational. Say $y_n = p_n/q_n$. Show that $\lim_{n \rightarrow \infty} q_n = \infty$. Now consider the function

$$f(x) \equiv \begin{cases} 0 & \text{if } x \text{ is irrational} \\ \frac{1}{q} & \text{if } x = \frac{p}{q} \text{ where the fraction is in lowest terms} \end{cases}$$

Show that f is continuous at each irrational number and discontinuous at every nonzero rational number. **Hint:** You ought to show that if $\frac{p_n}{q_n}$ is a sequence of rational numbers, p_n, q_n both integers converging to r an irrational number, then $\lim_{n \rightarrow \infty} q_n = \infty$. If it is not so, then $\{q_n\}$ would lie in some interval $[-m, m]$ and so there must be some integer k in this interval such that $q_n = k$ for infinitely many n . Now consider a subsequence n_j such that $q_{n_j} = k$ for all j . Argue that for j large enough, p_{n_j} must be constant and conclude that r must be rational after all.

11. Use the sequential definition of continuity described above to give an easy proof of Theorem 4.0.6.

12. Let $f(x) = \sqrt{x}$ show f is continuous at every value of x in its domain. For now, assume \sqrt{x} exists for all positive x . **Hint:** You might want to make use of the identity,

$$\sqrt{x} - \sqrt{y} = \frac{x - y}{\sqrt{x} + \sqrt{y}}$$

at some point in your argument.

13. Using Theorem 4.0.6, show all polynomials are continuous and that a rational function is continuous at every point of its domain. **Hint:** First show the function given as $f(x) = x$ is continuous and then use the Theorem 4.0.6. What about the case where x can be in \mathbb{R} ? Does the same conclusion hold?
14. Let $f(x) = \begin{cases} 1 & \text{if } x \in \mathbb{Q} \\ 0 & \text{if } x \notin \mathbb{Q} \end{cases}$ and consider $g(x) = f(x)(x - x^3)$. Determine where g is continuous and explain your answer.
15. Suppose f is any function whose domain is the integers. Thus $D(f) = \mathbb{Z}$, the set of whole numbers, $\dots, -3, -2, -1, 0, 1, 2, 3, \dots$. Then f is continuous. Why? **Hint:** In the definition of continuity, what if you let $\delta = \frac{1}{4}$? Would this δ work for a given $\varepsilon > 0$? This shows that the idea that a continuous function is one for which you can draw the graph without taking the pencil off the paper is a lot of nonsense.
16. Give an example of a function f which is not continuous at some point but $|f|$ is continuous at that point.
17. Find two functions which fail to be continuous but whose product is continuous.
18. Find two functions which fail to be continuous but whose sum is continuous.
19. Find two functions which fail to be continuous but whose quotient is continuous.
20. Suppose f is a function defined on \mathbb{R} and f is continuous at 0. Suppose also that $f(x+y) = f(x) + f(y)$. Show that if this is so, then f must be continuous at every value of $x \in \mathbb{R}$. Next show that for every rational number, r , $f(r) = rf(1)$. Finally explain why $f(r) = rf(1)$ for every r a real number. **Hint:** To do this last part, you need to use the density of the rational numbers and continuity of f .

4.3 The Extreme Values Theorem

The extreme values theorem says continuous functions achieve their maximum and minimum provided they are defined on a sequentially compact set. It was done by Bolzano in the 1830's and later by Weierstrass. This is a very significant theorem. It depends on continuity and on the function being defined on a compact set.

Example 4.3.1 Let $f(x) = 1/x$ for $x \in (0, 1)$.

Clearly, f is not bounded so it has no maximum although f is indeed continuous on this interval. The problem is that $(0, 1)$ is not compact. The same function defined on $[.000001, 1)$ would achieve its maximum but not its minimum. The following is the extreme value theorem or max, min theorem. I am being vague about where or what K is because it tends not to matter as long as it is compact.

Theorem 4.3.2 *Let K be sequentially compact and let $f : K \rightarrow \mathbb{R}$ be continuous. Then f achieves its maximum and its minimum on K . This means there exist, $x_1, x_2 \in K$ such that for all $x \in K$, $f(x_1) \leq f(x) \leq f(x_2)$.*

Proof: Let $M \equiv \sup \{f(x) : x \in K\}$. From the definition of the supremum, there exists $f(x_n)$ such that $\lim_{n \rightarrow \infty} f(x_n) = M$. This is because if $l < M$, there must be some x such that $f(x) \in (l, M]$ since otherwise M is not as defined. By sequential compactness, there is a subsequence $\{x_{n_k}\}$ such that $\lim_{k \rightarrow \infty} x_{n_k} = x \in K$. Then by continuity, $f(x) = \lim_{k \rightarrow \infty} f(x_{n_k}) = M$. That f achieves its minimum is proved exactly the same way. In particular, this shows that every function continuous on a sequentially compact set is bounded. ■

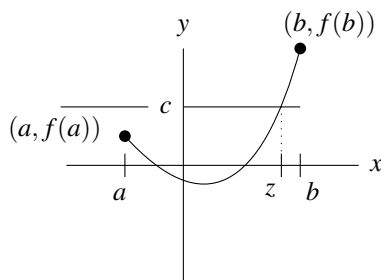
In fact a continuous function takes compact sets to compact sets. This is another of those big theorems which tends to hold whenever it makes sense. Therefore, I will be vague about the domain and range of the function f .

Theorem 4.3.3 *Let $D(f) \supseteq K$ where K is a sequentially compact set. Then $f(K)$ is also sequentially compact.*

Proof: Let $\{f(k_n)\}$ be a sequence in $f(K)$ so $\{k_n\}$ is a sequence in K . Since K is sequentially compact, there is a subsequence $\{k_{n_j}\}$ such that $\lim_{j \rightarrow \infty} k_{n_j} = k \in K$. By continuity, $\lim_{j \rightarrow \infty} f(k_{n_j}) = f(k) \in f(K)$. Thus $f(K)$ is sequentially compact as claimed. ■

4.4 The Intermediate Value Theorem

The next big theorem is called the intermediate value theorem and the following picture illustrates its conclusion.



It gives the existence of a certain point. You see in the picture there is a horizontal line, $y = c$ and a continuous function which starts off less than c at the point a and ends up greater than c at point b . The intermediate value theorem says there is some point between a and b shown in the picture as z such that the value of the function at this point equals c .

The theorem is due to Bolzano in 1817. You might think that this is an obvious theorem but this is not the case. It is not even true if you only had the rational numbers and this includes all numbers we work with. Nor is it true if you consider solutions to polynomial equations as was the case with issues related to the fundamental theorem of algebra. Consider rational numbers. Then $f(x) = x^2 - 2$ is continuous. $f(0) < 0$ and $f(2) > 0$ but the only point between 0 and 2 where this function is 0 is the point $\sqrt{2}$ which has been known for thousands of years to be irrational. You have to use something which rules out holes in

the real line. Bolzano's major contribution was to identify the concept of completeness as the reason for this theorem rather than some sort of vague notion based on pictures like the one just drawn. He did this long before Dedekind gave a way to construct the real numbers.

Proposition 4.4.1 *Let $f : [a, b] \rightarrow \mathbb{R}$ be continuous and suppose $f(a)f(b) \leq 0$. Then there exists $x \in [a, b]$ such that $f(x) = 0$.*

Proof: When we have an interval $[a_n, b_n]$ in this argument, c_n will be the midpoint $(a_n + b_n)/2$. Let $a_0 = a, b_0 = b$. If $[a_n, b_n]$ has been chosen such that $f(a_n)f(b_n) \leq 0$, consider $[a_n, c_n]$ and $[c_n, b_n]$. Either $f(a_n)f(c_n) \leq 0$ or $f(c_n)f(b_n) \leq 0$ since if both products are positive, then

$$f(a_n)f(c_n)f(c_n)f(b_n) = f(c_n)^2 f(a_n)f(b_n) > 0$$

so $f(a_n)f(b_n) > 0$. Pick one of the intervals for which the product is non-positive. Let the left endpoint be a_{n+1} and the right endpoint be b_{n+1} so $f(a_{n+1})f(b_{n+1}) \leq 0$. Continue picking the correct subinterval. These nested intervals have exactly one point in their intersection because they have diameters converging to 0. Call it x . Then

$$(f(x))^2 = \lim_{n \rightarrow \infty} f(a_n)f(b_n) \leq 0$$

This is by Theorem 4.0.7. Thus $f(x) = 0$. ■

It is easy to generalize this Proposition.

Theorem 4.4.2 *Suppose $f : [a, b] \rightarrow \mathbb{R}$ is continuous and suppose either $f(a) < c < f(b)$ or $f(a) > c > f(b)$. Then there exists $x \in (a, b)$ such that $f(x) = c$.*

Proof: Apply the above proposition to $g(x) \equiv f(x) - c$ obtaining a point $x \in (a, b)$ with $g(x) = f(x) - c = 0$. ■

Here is another lemma which may seem obvious but when you ask why, you begin to see that it is not as obvious as you thought. In fact, this is a special case of a general theory which says that one to one continuous functions from U , an open set in \mathbb{R}^p to \mathbb{R}^p take open sets to open sets. This is a very difficult result. The notation 1 – 1 means one to one. That is, if $x \neq y$, then $f(x) \neq f(y)$.

Lemma 4.4.3 *Let $\phi : [a, b] \rightarrow \mathbb{R}$ be a continuous function and suppose ϕ is 1 – 1 on (a, b) . Then ϕ is either strictly increasing or strictly decreasing on $[a, b]$.*

Proof: First it is shown that ϕ is either strictly increasing or strictly decreasing on (a, b) .

If ϕ is not strictly decreasing on (a, b) , then there exists $x_1 < y_1, x_1, y_1 \in (a, b)$ such that $(\phi(y_1) - \phi(x_1))(y_1 - x_1) > 0$. If for some other pair of points, $x_2 < y_2$ with $x_2, y_2 \in (a, b)$, the above inequality does not hold, then since ϕ is 1 – 1, $(\phi(y_2) - \phi(x_2))(y_2 - x_2) < 0$. Let $x_t \equiv tx_1 + (1 - t)x_2$ and $y_t \equiv ty_1 + (1 - t)y_2$. Then $x_t < y_t$ for all $t \in [0, 1]$ because

$$tx_1 \leq ty_1 \text{ and } (1 - t)x_2 \leq (1 - t)y_2$$

with strict inequality holding for at least one of these inequalities since not both t and $(1 - t)$ can equal zero. Now define

$$h(t) \equiv (\phi(y_t) - \phi(x_t))(y_t - x_t).$$

Since h is continuous and $h(0) < 0$, while $h(1) > 0$, there exists $t \in (0, 1)$ such that $h(t) = 0$. Therefore, both x_t and y_t are points of (a, b) and $\phi(y_t) - \phi(x_t) = 0$ contradicting the assumption that ϕ is one to one. It follows ϕ is either strictly increasing or strictly decreasing on (a, b) .

This property of being either strictly increasing or strictly decreasing on (a, b) carries over to $[a, b]$ by the continuity of ϕ . Suppose ϕ is strictly increasing on (a, b) . (A similar argument holds for ϕ strictly decreasing on (a, b) .) If $x > a$, then let z_n be a decreasing sequence of points of (a, x) converging to a . Then by continuity of ϕ at a ,

$$\phi(a) = \lim_{n \rightarrow \infty} \phi(z_n) \leq \phi(z_1) < \phi(x).$$

Therefore, $\phi(a) < \phi(x)$ whenever $x \in (a, b)$. Similarly $\phi(b) > \phi(x)$ for all $x \in (a, b)$. ■

4.5 Continuity of the Inverse

The inverse of a continuous function defined on an open interval is also continuous. This is an amazing result.

Corollary 4.5.1 *Let $f : (a, b) \rightarrow \mathbb{R}$ be one to one and continuous. Then $f(a, b)$ is an open interval, (c, d) and $f^{-1} : (c, d) \rightarrow (a, b)$ is continuous. If $f : [a, b] \rightarrow \mathbb{R}$ is one to one and continuous, then f^{-1} is also continuous.*

Proof: Consider the first part. By Lemma 4.4.3, f is strictly increasing or strictly decreasing. Hence $(f^{-1})^{-1}((x, y)) \equiv f((x, y))$ is an open interval. By Theorem 4.1.1, f^{-1} is continuous because inverse images of open intervals are open intervals.

As to the second claim, here is a direct proof based on notions of compactness. Say $f(x_n) \rightarrow f(x)$ where each x_n, x are in $[a, b]$. Does $x_n \rightarrow x$? If not, there exists a subsequence $\{x_{n_k}\}$ and $\varepsilon > 0$ such that $|x_{n_k} - x| \geq \varepsilon > 0$. However, by compactness, there is a further subsequence, $\{x_{n_{k_l}}\}$ such that $\lim_{l \rightarrow \infty} x_{n_{k_l}} = \hat{x}$ and so, by continuity,

$$f(\hat{x}) = \lim_{l \rightarrow \infty} f(x_{n_{k_l}}) = \lim_{n \rightarrow \infty} f(x_n) = f(x).$$

But $|\hat{x} - x| = \lim_{l \rightarrow \infty} |x_{n_{k_l}} - x| \geq \varepsilon$ and so this violates the assumption that f is one to one. Hence, $x_n \rightarrow x$ and so $f^{-1}(f(x_n)) \rightarrow f^{-1}(f(x))$ and so f^{-1} is continuous at every $f(x) \in f([a, b])$. ■

4.6 Exercises

1. Give an example of a continuous function defined on $(0, 1)$ which does not achieve its maximum on $(0, 1)$.
2. Give an example of a continuous function defined on $(0, 1)$ which is bounded but which does not achieve either its maximum or its minimum.
3. Give an example of a discontinuous function defined on $[0, 1]$ which is bounded but does not achieve either its maximum or its minimum.

4. Give an example of a continuous function defined on $[0, 1) \cup (1, 2]$ which is positive at 2, negative at 0 but is not equal to zero for any value of x .
5. Let $f(x) = x^5 + ax^4 + bx^3 + cx^2 + dx + e$ where a, b, c, d , and e are numbers. Show there exists real x such that $f(x) = 0$.
6. Give an example of a function which is one to one but neither strictly increasing nor strictly decreasing.
7. Show that the function $f(x) = x^n - a$, where n is a positive integer and a is a number, is continuous.
8. Use the intermediate value theorem on the function $f(x) = x^7 - 8$ to show $\sqrt[7]{8}$ must exist. State and prove a general theorem about n^{th} roots of positive numbers.
9. Prove $\sqrt{2}$ is irrational. **Hint:** Suppose $\sqrt{2} = p/q$ where p, q are positive integers and the fraction is in lowest terms. Then $2q^2 = p^2$ and so p^2 is even. Explain why $p = 2r$ so p must be even. Next argue q must be even.
10. Let $f(x) = x - \sqrt{2}$ for $x \in \mathbb{Q}$, the rational numbers. Show that even though $f(0) < 0$ and $f(2) > 0$, there is no point in \mathbb{Q} where $f(x) = 0$. Does this contradict the intermediate value theorem? Explain.
11. A circular hula hoop lies partly in the shade and partly in the hot sun. Show there exist two points on the hula hoop which are at opposite sides of the hoop which have the same temperature. **Hint:** Imagine this is a circle and points are located by specifying their angle, θ from a fixed diameter. Then letting $T(\theta)$ be the temperature in the hoop, $T(\theta + 2\pi) = T(\theta)$. You need to have $T(\theta) = T(\theta + \pi)$ for some θ . Assume T is a continuous function of θ .
12. A car starts off on a long trip with a full tank of gas which is driven till it runs out of gas. Show that at some time the number of miles the car has gone exactly equals the number of gallons of gas in the tank.
13. Suppose f is a continuous function defined on $[0, 1]$ which maps $[0, 1]$ into $[0, 1]$. Show there exists $x \in [0, 1]$ such that $x = f(x)$. **Hint:** Consider $h(x) \equiv x - f(x)$ and the intermediate value theorem. This is a one dimensional version of the Brouwer fixed point theorem.
14. Let f be a continuous function on $[0, 1]$ such that $f(0) = f(1)$. Let n be a positive integer larger than 2. Show there must exist $c \in [0, 1 - \frac{1}{n}]$ such that $f(c + \frac{1}{n}) = f(c)$. **Hint:** Consider $h(x) \equiv f(x + \frac{1}{n}) - f(x)$. Consider the subintervals $[\frac{k-1}{n}, \frac{k}{n}]$ for $k = 1, \dots, n-1$. You want to show that h equals zero on one of these intervals. If h changes sign between two successive intervals, then you are done. Assume then, that this does not happen. Say h remains positive. Argue that $f(0) < f(\frac{n-1}{n})$. Thus $f(\frac{n-1}{n}) > f(1) = f(\frac{n-1}{n} + \frac{1}{n})$. It follows that $h(1 - \frac{1}{n}) < 0$ but $h(1 - \frac{2}{n}) > 0$.

4.7 Uniform Continuity

There is a theorem about the integral of a continuous function which requires the notion of uniform continuity. This is discussed in this section. Consider the function $f(x) = \frac{1}{x}$

for $x \in (0, 1)$. This is a continuous function because, by Theorem 4.0.6, it is continuous at every point of $(0, 1)$. However, for a given $\varepsilon > 0$, the δ needed in the ε, δ definition of continuity becomes very small as x gets close to 0. The notion of uniform continuity involves being able to choose a single δ which works on the whole domain of f . It is usually assumed that this concept belongs to the latter half of the nineteenth century and is due to Weierstrass because he used it most systematically, but it may have been understood by Cauchy and Bolzano.

Definition 4.7.1 *Let f be a function. Then f is uniformly continuous if for every $\varepsilon > 0$, there exists a δ depending only on ε such that for $x, y \in D(f)$, if $|x - y| < \delta$ then $|f(x) - f(y)| < \varepsilon$.*

It is an amazing fact that under certain conditions continuity implies uniform continuity. Recall that it was shown above that closed intervals are sequentially compact.

Theorem 4.7.2 *Let $f : K \rightarrow \mathbb{R}$ be continuous where K is a sequentially compact set in \mathbb{R} . Then f is uniformly continuous on K .*

Proof: If f is not uniformly continuous, there exists $\varepsilon > 0$ such that for every $\delta > 0$ there exists a pair of points, x_δ and y_δ such that even though $|x_\delta - y_\delta| < \delta$, it is the case that $|f(x_\delta) - f(y_\delta)| \geq \varepsilon$. Taking a succession of values for δ equal to $1, 1/2, 1/3, \dots$, and letting the exceptional pair of points for $\delta = 1/n$ be denoted by x_n and y_n ,

$$|x_n - y_n| < \frac{1}{n}, |f(x_n) - f(y_n)| \geq \varepsilon.$$

Now since K is sequentially compact, there exists a subsequence, $\{x_{n_k}\}$ such that $x_{n_k} \rightarrow z \in K$. Now $n_k \geq k$ and so $|x_{n_k} - y_{n_k}| < \frac{1}{k}$. Consequently, $y_{n_k} \rightarrow z$ also. (x_{n_k} is like a person walking toward a certain point and y_{n_k} is like a dog on a leash which is constantly getting shorter. Obviously y_{n_k} must move toward the point also. You should give a precise proof of what is needed here.) By continuity of f and Theorem 4.0.7, $0 = |f(z) - f(z)| = \lim_{k \rightarrow \infty} |f(x_{n_k}) - f(y_{n_k})| \geq \varepsilon$, an obvious contradiction. Therefore, the conclusion of this theorem must be true since it cannot be false. ■

4.8 Examples of Continuous Functions

Polynomials are continuous. Suppose $p(x) = a_n x^n + \dots + a_1 x$ is a polynomial. Then if $\lim_{k \rightarrow \infty} x_k = x$, it follows from properties of limits of sequences that $\lim_{k \rightarrow \infty} p(x_k) = p(x)$. See Theorem 3.3.7. Rational functions are continuous wherever the denominator is not zero by Theorem 4.0.6.

The sine function is continuous. To see this, suppose $\lim_{k \rightarrow \infty} x_k = x$

$$\begin{aligned} |\sin(x_k) - \sin(x)| &= |\sin(x + x_k - x) - \sin(x)| \\ &= |\sin(x) \cos(x_k - x) + \sin(x_k - x) \cos(x) - \sin(x)| \\ &\leq |\sin(x)| |\cos(x_k - x) - 1| + |\cos(x)| |\sin(x_k - x)| \\ &\leq |\cos(x_k - x) - 1| + |\sin(x_k - x)| \end{aligned}$$

Now, from the geometric definition of the sine and cosine, both terms on the right converge to 0 as $k \rightarrow \infty$. Thus the sine is continuous. Similar reasoning shows that the cosine is continuous.

Of course one can also consider the arcsin function defined as $\arcsin(y)$ is the angle whose sine is y which is in $[-\frac{\pi}{2}, \frac{\pi}{2}]$. This is a well defined function because the sine function is one to one on $[-\frac{\pi}{2}, \frac{\pi}{2}]$. By Corollary 4.5.1 this inverse function is continuous. The arccos function is defined on $[-1, 1]$ and $\arccos(y)$ is defined as the angle in $[0, \pi]$ whose cosine is y . By the same corollary, this function is also continuous.

It was observed that a small change in x led to a small change in $\ln(x)$. Thus \ln is continuous. It follows from Corollary 4.5.1 that its inverse \exp is continuous. Then from the theorem about various combinations of continuous functions, Theorem 4.0.6, all of the functions \log_b for $b \neq 1$, and all functions $x \rightarrow b^x$ for $b > 0$ are continuous.

As noted above, all polynomials are continuous as are all rational functions at all points of their domain. Indeed, if $p(x)/q(x)$ is a rational function and $q(x_0) \neq 0$, then if $x_n \rightarrow x_0$, it follows that $q(x_n) \rightarrow q(x_0) \neq 0$ and $p(x_n) \rightarrow p(x_0)$. Then from the theorem on limits of sequences, Theorem 3.3.7, it follows $p(x_n)/q(x_n) \rightarrow p(x_0)/q(x_0)$.

It is now clear that we have a very large collection of functions which are known to be continuous. The next chapter will consider something even better.

4.9 Sequences of Functions

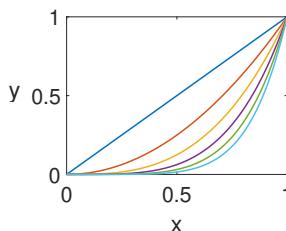
Suppose for each $n \in \mathbb{N}$, f_n is a continuous function defined on some interval $[a, b]$. Also suppose that for each fixed $x \in [a, b]$, $\lim_{n \rightarrow \infty} f_n(x) = f(x)$. This is called pointwise convergence. Does it follow that f is continuous on $[a, b]$? The answer is NO. Consider the following

$$f_n(x) \equiv x^n \text{ for } x \in [0, 1]$$

Then $\lim_{n \rightarrow \infty} f_n(x)$ exists for each $x \in [0, 1]$ and equals

$$f(x) \equiv \begin{cases} 1 & \text{if } x = 1 \\ 0 & \text{if } x \neq 1 \end{cases}$$

You should verify this is the case. This limit function is not continuous. Indeed, it has a jump at $x = 1$. Here are graphs of the first few of these functions.



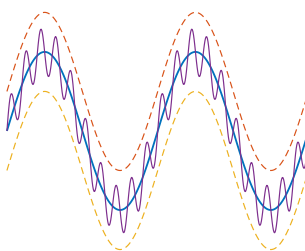
If you want the convergence to carry continuity with it you need something more than point-wise convergence. You need uniform convergence. The concept may have been understood by Cauchy but was not at all clear. Weierstrass is the first to formalize this concept and prove theorems like what follow.

Definition 4.9.1 Let $\{f_n\}$ be a sequence of functions defined on D . Then f_n is said to converge uniformly to f on D if

$$\lim_{n \rightarrow \infty} \|f_n - f\|_{\infty} \equiv \lim_{n \rightarrow \infty} \left(\sup_{x \in D} |f_n(x) - f(x)| \right) = 0$$

$\|\cdot\|_{\infty}$ is called a norm. It is defined on the functions f which are uniformly bounded, meaning $\|f\|_{\infty} \equiv \sup\{|f(x)| : x \in D\} < \infty$.

The following picture illustrates the above definition.



The dashed lines define a small tube centered about the graph of f and the graph of the function f_n fits in this tube for all n sufficiently large. In the picture, the function f is being approximated by f_n which is very wriggly.

It is convenient to observe the following properties of $\|\cdot\|_{\infty}$, written $\|\cdot\|$ for short.

Lemma 4.9.2 The norm $\|\cdot\|_{\infty}$ satisfies the following properties.

$$\|f\| \geq 0 \text{ and equals 0 if and only if } f = 0 \quad (4.1)$$

For α a number,

$$\|\alpha f\| = |\alpha| \|f\| \quad (4.2)$$

$$\|f + g\| \leq \|f\| + \|g\| \quad (4.3)$$

Proof: The first claim 4.1 is obvious. As to 4.2, it follows fairly easily.

$$\|\alpha f\| \equiv \sup_{x \in D} |\alpha f(x)| = \sup_{x \in D} |\alpha| |f(x)| = |\alpha| \sup_{x \in D} |f(x)| = |\alpha| \|f\|$$

The last follows from $|f(x) + g(x)| \leq |f(x)| + |g(x)| \leq \|f\| + \|g\|$. Therefore,

$$\sup_{x \in D} |f(x) + g(x)| \equiv \|f + g\| \leq \|f\| + \|g\| \quad \blacksquare$$

Now with this preparation, here is the main result. Again, I am being vague about the domain of the functions. This is because it does not matter much and it is a bad idea to get hung up on trivialities which don't matter.

Theorem 4.9.3 Let f_n be continuous and each f_n bounded on D and suppose that $\lim_{n \rightarrow \infty} \|f_n - f\| = 0$. Then f is also continuous. If each f_n is uniformly continuous, then f is uniformly continuous.

Proof: Let $\varepsilon > 0$ be given and let $x \in D$. Let n be such that $\|f_n - f\| < \frac{\varepsilon}{3}$. By continuity of f_n there exists $\delta > 0$ such that if $|y - x| < \delta$, then $|f_n(y) - f_n(x)| < \frac{\varepsilon}{3}$. Then for such y ,

$$\begin{aligned} |f(y) - f(x)| &\leq |f(y) - f_n(y)| + |f_n(y) - f_n(x)| + |f_n(x) - f(x)| \\ &< \frac{\varepsilon}{3} + \|f - f_n\| + \|f_n - f\| < \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon \end{aligned}$$

and so this shows that f is continuous. To show the claim about uniform continuity, use the same string of inequalities above where δ is chosen so that for any pair x, y with $|x - y| < \delta$, $|f_n(y) - f_n(x)| < \frac{\varepsilon}{3}$. Then the above shows that if $|x - y| < \delta$, then $|f(x) - f(y)| < \varepsilon$ which satisfies the definition of uniformly continuous. ■

This implies the following interesting corollary about a uniformly Cauchy sequence of continuous functions.

Definition 4.9.4 Let $\{f_n\}$ be a sequence of continuous functions defined on $[a, b]$. It is said to be uniformly Cauchy if for every $\varepsilon > 0$ there exists n_ε such that if $m, k > n_\varepsilon$, then $\|f_m - f_k\| < \varepsilon$.

Corollary 4.9.5 Suppose $\{f_n\}$ is a uniformly Cauchy sequence of continuous uniformly bounded functions defined on D . Then there exists a unique continuous function f such that

$$\lim_{n \rightarrow \infty} \|f_n - f\| = 0$$

If each f_n is uniformly continuous, then so is f .

Proof: The hypothesis implies that $\{f_n(x)\}$ is a Cauchy sequence in \mathbb{R} for each x . Therefore, by completeness of \mathbb{R} , Theorem 3.7.3, this sequence converges for each x . Let $f(x) \equiv \lim_{n \rightarrow \infty} f_n(x)$. Then for $m > n$,

$$|f(x) - f_n(x)| \leq \sup_m |f_m(x) - f_n(x)| \leq \sup_m \|f_m - f_n\| < \varepsilon$$

provided n is sufficiently large. Since x is arbitrary, it follows that $\|f - f_n\| \leq \varepsilon$ for all n large enough which shows by definition that $\lim_{n \rightarrow \infty} \|f_n - f\| = 0$.

Now the continuity of f follows from Theorem 4.9.3 and if each f_n is uniformly continuous, then so is f . How many such functions f are there? There can be only one because $f(x)$ must equal the limit of $f_n(x)$. ■

4.10 Polynomials and Continuous Functions

It turns out that if f is a continuous real valued function defined on an interval, $[a, b]$ then there exists a sequence of polynomials, $\{p_n\}$ such that the sequence converges uniformly to f on $[a, b]$. I will first show this is true for the interval $[0, 1]$ and then verify it is true on any closed and bounded interval. First here is a little lemma which is interesting in probability. It is actually an estimate for the variance of a binomial distribution.

Lemma 4.10.1 The following estimate holds for $x \in [0, 1]$ and $n \geq 2$.

$$\sum_{k=0}^n \binom{n}{k} (k - nx)^2 x^k (1 - x)^{n-k} \leq \frac{1}{4}n$$

Proof: Here are some observations. $\sum_{k=0}^n \binom{n}{k} k x^k (1-x)^{n-k} =$

$$\begin{aligned} & nx \sum_{k=1}^n \frac{(n-1)!}{(k-1)!((n-1)-(k-1))!} x^{k-1} (1-x)^{(n-1)-(k-1)} \\ &= nx \sum_{k=0}^{n-1} \binom{n-1}{k} x^k (1-x)^{n-1-k} = nx \\ & \sum_{k=0}^n \binom{n}{k} k(k-1) x^k (1-x)^{n-k} \\ &= n(n-1) x^2 \sum_{k=2}^n \frac{(n-2)!}{(k-2)!(n-2-(k-2))!} x^{k-2} (1-x)^{(n-2)-(k-2)} \\ &= n(n-1) x^2 \sum_{k=0}^{n-2} \binom{n-2}{k} x^k (1-x)^{(n-2)-k} = n(n-1) x^2 \end{aligned}$$

Now $(k-nx)^2 = k^2 - 2knx + n^2x^2 = k(k-1) + k(1-2nx) + n^2x^2$. From the above and the binomial theorem, $\sum_{k=0}^n \binom{n}{k} (k-nx)^2 x^k (1-x)^{n-k} =$

$$\begin{aligned} & \sum_{k=0}^n \binom{n}{k} k(k-1) x^k (1-x)^{n-k} + (1-2nx) \sum_{k=0}^n \binom{n}{k} k x^k (1-x)^{n-k} \\ &+ n^2 x^2 \sum_{k=0}^n \binom{n}{k} x^k (1-x)^{n-k} = n(n-1) x^2 + (1-2nx) nx + n^2 x^2 \\ &= nx(1-x) \leq n \frac{1}{4} \blacksquare \end{aligned}$$

Now let f be a continuous function defined on $[0, 1]$. Let p_n be the polynomial defined by

$$p_n(x) \equiv \sum_{k=0}^n \binom{n}{k} f\left(\frac{k}{n}\right) x^k (1-x)^{n-k}. \quad (4.4)$$

Theorem 4.10.2 *The sequence of polynomials in 4.4 converges uniformly to f on $[0, 1]$. These polynomials are called the Bernstein polynomials.*

Proof: By the binomial theorem,

$$f(x) = f(x) \sum_{k=0}^n \binom{n}{k} x^k (1-x)^{n-k} = \sum_{k=0}^n \binom{n}{k} f(x) x^k (1-x)^{n-k}$$

and so by the triangle inequality

$$|f(x) - p_n(x)| \leq \sum_{k=0}^n \binom{n}{k} \left| f\left(\frac{k}{n}\right) - f(x) \right| x^k (1-x)^{n-k}$$

. By Theorems 3.6.2 and 4.7.2, f is uniformly continuous. Let δ go with $\varepsilon/2$ in the definition of uniform continuity. At this point you break the sum into two pieces, those values

of k such that $|k/n - x| < \delta$ and those values of k where $|x - (k/n)| \geq \delta$. Then from the Lemma ??,

$$\begin{aligned}
 |f(x) - p_n(x)| &\leq \sum_{|x - (k/n)| < \delta} \binom{n}{k} \left| f\left(\frac{k}{n}\right) - f(x) \right| x^k (1-x)^{n-k} \\
 &\quad + \sum_{|x - (k/n)| \geq \delta} \binom{n}{k} \left| f\left(\frac{k}{n}\right) - f(x) \right| x^k (1-x)^{n-k} \quad (4.5) \\
 &\leq \sum_{|x - (k/n)| < \delta} \binom{n}{k} \frac{\varepsilon}{2} x^k (1-x)^{n-k} + 2M \sum_{|nx - k| \geq n\delta} \binom{n}{k} x^k (1-x)^{n-k} \\
 &\leq \frac{\varepsilon}{2} \sum_{k=0}^n \binom{n}{k} x^k (1-x)^{n-k} + 2M \sum_{|nx - k| \geq n\delta} \binom{n}{k} \frac{(k - nx)^2}{n^2 \delta^2} x^k (1-x)^{n-k} \\
 &\leq \frac{\varepsilon}{2} + 2M \frac{1}{4} n \frac{1}{n^2 \delta^2} = \frac{\varepsilon}{2} + \frac{1}{2} \frac{M}{n \delta^2}
 \end{aligned}$$

Therefore, whenever n is sufficiently large that $\frac{4M}{n\delta^2} < \frac{\varepsilon}{2}$, it follows that for all n this large and $x \in [0, 1]$,

$$|f(x) - p_n(x)| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \blacksquare$$

Now this theorem has been done, it is easy to extend to continuous functions defined on $[a, b]$. This yields the celebrated Weierstrass approximation theorem.

Theorem 4.10.3 Suppose f is a continuous function defined on $[a, b]$. Then there exists a sequence of polynomials, $\{p_n\}$ which converges uniformly to f on $[a, b]$.

Proof: For $t \in [0, 1]$, let $h(t) = a + (b - a)t$. Thus h maps $[0, 1]$ one to one and onto $[a, b]$. Thus $f \circ h$ is a continuous function defined on $[0, 1]$. It follows there exists a sequence of polynomials $\{p_n\}$ defined on $[0, 1]$ which converges uniformly to $f \circ h$ on $[0, 1]$. Thus for every $\varepsilon > 0$ there exists N_ε such that if $n \geq N_\varepsilon$, then for all $t \in [0, 1]$,

$$|f \circ h(t) - p_n(t)| < \varepsilon.$$

However, h is onto and one to one and so for all $x \in [a, b]$, $|f(x) - p_n(h^{-1}(x))| < \varepsilon$. Now note that the function $x \rightarrow p_n(h^{-1}(x))$ is a polynomial because $h^{-1}(x) = \frac{x-a}{b-a}$. More specifically, if $p_n(t) = \sum_{k=0}^m a_k t^k$ it follows

$$p_n(h^{-1}(x)) = \sum_{k=0}^m a_k \left(\frac{x-a}{b-a} \right)^k$$

which is clearly another polynomial. \blacksquare

Weierstrass did not prove this theorem in this way. He used integrals instead of sums to do it, but integrals have not been discussed yet. I think the Bernstein polynomials used here give the easiest proof. This amazing theorem shows that every continuous function defined on a finite closed interval is the uniform limit of polynomials. The analog does not hold for continuous functions of complex variables but this is another topic entirely.

4.11 Exercises

1. A function f is Lipschitz continuous or just Lipschitz for short if there exists a constant, K such that

$$|f(x) - f(y)| \leq K|x - y|$$

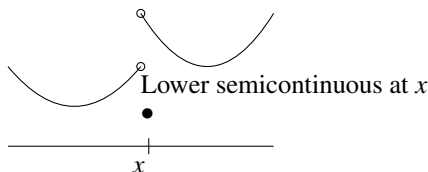
for all $x, y \in D$. Show every Lipschitz function is uniformly continuous.

2. If $|x_n - y_n| \rightarrow 0$ and $x_n \rightarrow z$, show that $y_n \rightarrow z$ also. This was used in the proof of Theorem 4.7.2.
3. Consider $f : (1, \infty) \rightarrow \mathbb{R}$ given by $f(x) = \frac{1}{x}$. Show f is uniformly continuous even though the set on which f is defined is not sequentially compact.
4. If f is uniformly continuous, does it follow that $|f|$ is also uniformly continuous? If $|f|$ is uniformly continuous does it follow that f is uniformly continuous? Answer the same questions with “uniformly continuous” replaced with “continuous”. Explain why.
5. Suppose f is a continuous function defined on D and $\lambda \equiv \inf\{f(x) : x \in D\}$. A sequence $\{x_n\}$ of points of D is called a minimizing sequence if

$$\lim_{n \rightarrow \infty} f(x_n) = \lambda.$$

A maximizing sequence is defined analogously. Show that minimizing sequences and maximizing sequences always exist. Now let K be a sequentially compact set and $f : K \rightarrow \mathbb{R}$. Show that f achieves both its maximum and its minimum on K by considering directly minimizing and maximizing sequences. **Hint:** Let $M \equiv \sup\{f(x) : x \in K\}$. Argue there exists a sequence, $\{x_n\} \subseteq K$ such that $f(x_n) \rightarrow M$. Now use sequential compactness to get a subsequence, $\{x_{n_k}\}$ such that $\lim_{k \rightarrow \infty} x_{n_k} = x \in K$ and use the continuity of f to verify that $f(x) = M$. Incidentally, this shows f is bounded on K as well. A similar argument works to give the part about achieving the minimum.

6. Let $f : D \rightarrow \mathbb{R}$ be a function. This function is said to be lower semicontinuous³



at $x \in D$ if for any sequence $\{x_n\} \subseteq D$ which converges to x it follows $f(x) \leq \liminf_{n \rightarrow \infty} f(x_n)$. Suppose D is sequentially compact and f is lower semicontinuous at every point of D . Show that then f achieves its minimum on D .

7. Let $f : D \rightarrow \mathbb{R}$ 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 $f(x) \geq \limsup_{n \rightarrow \infty} f(x_n)$. Suppose D is sequentially compact and f is upper semicontinuous at every point of D . Show that then f achieves its maximum on D .

³The notion of lower semicontinuity is very important for functions which are defined on infinite dimensional sets. In more general settings, one formulates the concept differently.

8. Show that a real valued function is continuous if and only if it is both upper and lower semicontinuous.
9. Give an example of a lower semicontinuous function which is not continuous and an example of an upper semicontinuous function which is not continuous.
10. Suppose $\{f_\alpha : \alpha \in \Lambda\}$ is a collection of continuous functions. Define the function $F(x) \equiv \inf\{f_\alpha(x) : \alpha \in \Lambda\}$. 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.
11. Let f be a function. $\text{epi}(f)$ is defined as $\{(x, y) : y \geq f(x)\}$. It is called the epigraph of f . We say $\text{epi}(f)$ is closed if whenever $(x_n, y_n) \in \text{epi}(f)$ and $x_n \rightarrow x$ and $y_n \rightarrow y$, it follows $(x, y) \in \text{epi}(f)$. Show f is lower semicontinuous if and only if $\text{epi}(f)$ is closed. What would be the corresponding result equivalent to upper semicontinuous?
12. Explain why $x \rightarrow \exp(\sin(\ln(x^2 + 1)))$ is continuous on \mathbb{R} .
13. Suppose $f : \mathbb{N} \rightarrow \mathbb{R}$ is a function. Here \mathbb{N} is the set of positive integers. Explain why f is continuous. Is it necessarily uniformly continuous? Note that you cannot graph this function without taking pencil off the paper.

4.12 Limit of a Function

One of the main reasons for discussing limits of functions is to allow a definition of the derivative. Continuity, derivatives, and integrals are the three main topics in calculus. So far, all that has been discussed is continuity. The derivative will be in the next chapter.

In this section, functions will be defined on some nonempty subset of \mathbb{R} .

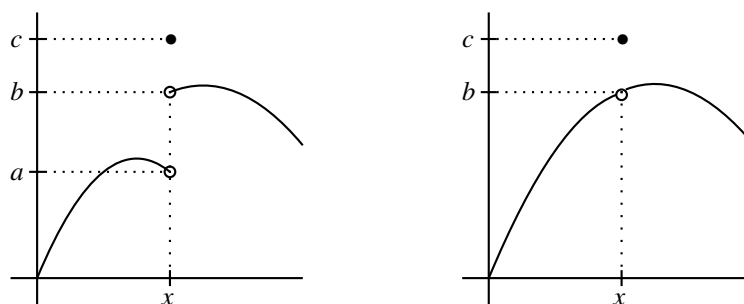
Definition 4.12.1 *A point x is a limit point of a nonempty set D means that $B(x, \delta)$ always contains a point of D different than x for any $\delta > 0$. Then if D is the domain of a function f , and $L \in \mathbb{R}$, we say that $\lim_{y \rightarrow x} f(y) = L$ means: For every $\varepsilon > 0$, there exists a $\delta > 0$ such that whenever $y \in D$ and $0 < |y - x| < \delta$, it follows that $|f(y) - L| < \varepsilon$.*

If x is a limit point of $D_+ \equiv D \cap (x, \infty)$, then $\lim_{y \rightarrow x^+} f(y) = L$ means the same thing except y is restricted to D_+ . If x is a limit point of $D_- \equiv D \cap (-\infty, x)$, then $\lim_{y \rightarrow x^-} f(y) = L$ means the same thing except you restrict y to D_- .

Limits are also taken as a variable “approaches” infinity. Of course nothing is “close” to infinity and so this requires a slightly different definition. Suppose D contains all x sufficiently large. Then $\lim_{x \rightarrow \infty} f(x) = L$ if for every $\varepsilon > 0$ there exists l such that whenever $x > l$, $|f(x) - L| < \varepsilon$ and $\lim_{x \rightarrow -\infty} f(x) = L$ if for every $\varepsilon > 0$ there exists l such that whenever $x < l$, $|f(x) - L| < \varepsilon$ holds.

The main example of interest in this book is when the limit point is either the interior of an interval, the end point of an interval or an end point of two adjacent intervals.

The following pictures illustrate some of these definitions.



In the left picture is shown the graph of a function. Note the value of the function at x equals c while $\lim_{y \rightarrow x+} f(y) = b$ and $\lim_{y \rightarrow x-} f(y) = a$. In the second picture, $\lim_{y \rightarrow x} f(y) = b$. Note that the value of the function at the point x has nothing to do with the limit of the function in any of these cases. **The value of a function at x has nothing to do with the value of the limit at x !** This must always be kept in mind. You do not evaluate interesting limits by computing $f(x)$! In the above picture, $f(x)$ is always wrong! It may be the case that $f(x)$ is right but this is merely a happy coincidence when it occurs and as explained below in Theorem 4.12.7, this is sometimes equivalent to f being continuous at x . Indeed, the concept of limit really only gives you something new and interesting when the function you are taking the limit of is not defined at the point. To repeat: **You do not evaluate interesting limits by plugging in a value!**

Theorem 4.12.2 *If $\lim_{y \rightarrow x} f(y) = L$ and $\lim_{y \rightarrow x} f(y) = L_1$, then $L = L_1$. The same conclusion follows in the case of $\lim_{y \rightarrow x+}$, $\lim_{y \rightarrow x-}$, $\lim_{y \rightarrow \infty}$, $\lim_{y \rightarrow -\infty}$.*

Proof: Let $\varepsilon > 0$ be given. There exists $\delta > 0$ such that if $0 < |y - x| < \delta$, then

$$|f(y) - L| < \varepsilon, |f(y) - L_1| < \varepsilon.$$

Therefore, for such y which exists because x is a limit point,

$$|L - L_1| \leq |L - f(y)| + |f(y) - L_1| < \varepsilon + \varepsilon = 2\varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, this shows $L = L_1$. The last claims are similar. For example, if it is $\lim_{y \rightarrow \infty}$, then there exists l such that if $y > l$ then

$$|f(y) - L| < \varepsilon, |f(y) - L_1| < \varepsilon$$

and then the same argument just used shows that $|L - L_1| < 2\varepsilon$. ■

Another concept is that of a function having either ∞ or $-\infty$ as a limit. In this case, the values of the function do not ever get close to their “limit” because nothing can be close to $\pm\infty$. Roughly speaking, the limit of the function equals ∞ if the values of the function are ultimately larger than any given number. More precisely:

Definition 4.12.3 *If $f(x) \in \mathbb{R}$, then $\lim_{y \rightarrow x} f(x) = \infty$ if for every number l , there exists $\delta > 0$ such that whenever $|y - x| < \delta$, then $f(x) > l$. $\lim_{x \rightarrow \infty} f(x) = \infty$ if for all k , there exists l such that $f(x) > k$ whenever $x > l$. One sided limits and limits as the variable approaches $-\infty$, are defined similarly.*

It may seem there is a lot to memorize here. In fact, this is not so because all the definitions are intuitive when you understand them. Everything becomes much easier when you understand the definitions. This is usually the way it works in mathematics.

In the following theorem it is assumed the domains of the functions are such that the various limits make sense. Thus, if $\lim_{y \rightarrow x}$ is used, it is to be understood the function is defined on $(x - \delta, x) \cup (x, x + \delta)$ for some $\delta > 0$. However, to avoid having to state things repetitively this symbol will be written to symbolize $\lim_{y \rightarrow x+}$ or $\lim_{y \rightarrow x-}$ and in either of these cases, it is understood the function is defined on an appropriate set so that the limits make sense. Thus in the case of $\lim_{y \rightarrow x+}$ the function is understood to be defined on an interval of the form $(x, x + \delta)$ with a similar convention holding for $\lim_{y \rightarrow x-}$.

To reduce to the consideration of sequences, here is a nice result.

Proposition 4.12.4 *Let x be a limit point of $D(f)$. Then $\lim_{y \rightarrow x} f(y) = L$ if and only if whenever $x_n \rightarrow x$ for each $x_n \neq x$, the x_n distinct points, it follows that $f(x_n) \rightarrow L$.*

Proof: \Rightarrow Let $x_n \rightarrow x$ where no x_n equals x . Let $\varepsilon > 0$ be given. By assumption, $|f(y) - L| < \varepsilon$ whenever $0 < |y - x| < \delta$ for some δ . However, for all n large enough, $0 < |x_n - x| < \delta$ and so $|f(x_n) - L| < \varepsilon$. Hence $f(x_n) \rightarrow L$.

\Leftarrow Suppose the condition on the sequences holds. If the condition for the limit does not hold, then there exists $\varepsilon > 0$ such that no matter how small δ , there will be $0 < |y - x| < \delta, y \in D(f)$, and yet $|f(y) - L| \geq \varepsilon$. Now let $\delta_1 = 1$. There exists $x_1 \neq x$ with $x_1 \in B(x, \delta_1) \cap D(f)$ and $|f(x_1) - L| \geq \varepsilon$. Let $\delta_2 \equiv \min\left(\frac{1}{2}, \frac{1}{2}|x - x_1|\right)$. Now pick $x_2 \in B(x, \delta_2), x_2 \neq x$ such that $|f(x_2) - L| \geq \varepsilon$. Let $\delta_3 \equiv \min\left(\frac{1}{2^3}, \frac{1}{2}|x - x_1|, \frac{1}{2}|x - x_2|\right)$ and pick $x_3 \in B(x, \delta_3)$ with $|f(x_3) - L| \geq \varepsilon, x_3 \neq x$. Continue this way to generate a sequence of distinct points $\{x_n\}$, none equal to x which converges to x . Then $L = \lim_{n \rightarrow \infty} f(x_n)$ because of the condition on limits of the sequence so eventually $|L - f(x_n)| < \varepsilon$, contrary to the construction of the x_n . ■

Theorem 4.12.5 *In this theorem, the symbol $\lim_{y \rightarrow x}$ denotes any of the limits described above. Suppose $\lim_{y \rightarrow x} f(y) = L$ and $\lim_{y \rightarrow x} g(y) = K$ where K and L are numbers, not $\pm\infty$. Then if a, b are numbers,*

$$\lim_{y \rightarrow x} (af(y) + bg(y)) = aL + bK, \quad (4.6)$$

$$\lim_{y \rightarrow x} fg(y) = LK \quad (4.7)$$

and if $K \neq 0$,

$$\lim_{y \rightarrow x} \frac{f(y)}{g(y)} = \frac{L}{K}. \quad (4.8)$$

Also, if h is a continuous function defined in some interval containing L , then

$$\lim_{y \rightarrow x} h \circ f(y) = h(L). \quad (4.9)$$

Suppose f is real valued and $\lim_{y \rightarrow x} f(y) = L$. If $f(y) \leq a$ all y near x either to the right or to the left of x , then $L \leq a$ and if $f(y) \geq a$ then $L \geq a$.

Proof: All of these follow from the limit theorem for sequences and Proposition 4.12.4. For example, consider 4.8. Let $x_n \rightarrow x$ the x_n distinct and none equal to x . Then $|g(x_n)| >$

$|K|/2$ for large enough n . Here is why: $|g(x_n) - K| < \frac{|K|}{2}$ if n large enough and so, by the triangle inequality, $|g(x_n)| > |K|/2$ for large enough n . Thus for such n ,

$$\left| \frac{f(x_n)}{g(x_n)} - \frac{L}{K} \right| \leq \left| \frac{Kf(x_n) - Lg(x_n)}{K(K/2)} \right| \leq \frac{2}{K^2} |Kf(x_n) - Lg(x_n)|$$

and the right side inside the absolute value converges to $KL - LK = 0$.

Consider 4.9. Let $x_n \rightarrow x$ where none of the $x_n = x$. Then $h \circ f(x_n) = h(f(x_n))$ and since $f(x_n) \rightarrow L$ and h is continuous near L , $h(f(x_n)) \rightarrow h(L)$. The other two claims are somewhat easier and follow from the same methods. ■

A very useful theorem for finding limits is called the squeezing theorem.

Theorem 4.12.6 Suppose f, g, h are real valued functions and that

$$\lim_{x \rightarrow a} f(x) = L = \lim_{x \rightarrow a} g(x)$$

and for all x near a ,

$$f(x) \leq h(x) \leq g(x). \quad (*)$$

Then

$$\lim_{x \rightarrow a} h(x) = L.$$

Proof: Let $\{x_n\}$ be a sequence with $\lim_{n \rightarrow \infty} x_n = a$ and the x_n are distinct. Then $f(x_n) \leq h(x_n) \leq g(x_n)$ for all n large enough, and by Theorem 3.3.15 about limits of sequences,

$$L = \lim_{n \rightarrow \infty} g(x_n) \geq \lim_{n \rightarrow \infty} h(x_n) \geq \lim_{n \rightarrow \infty} f(x_n) = L. \quad \blacksquare$$

Note that the end points of an interval are always limit points of the interval.

Next is the relation between limits and continuity. I am being vague about there f has its values and $D(f)$ because this is one of those things which is nearly always the case. Go ahead and make $D(f) \subseteq \mathbb{R}$ if you like but it won't end up mattering much.

Theorem 4.12.7 Let f be a function defined on $D(f)$. Then f is continuous at a limit point $x \in D(f)$ if and only if $\lim_{y \rightarrow x} f(y) = f(x)$.

Proof: \Rightarrow Suppose x_n is any sequence of distinct points of $D(f)$ which converges to the limit point x none of which equal x . Then by continuity, $f(x_n) \rightarrow f(x)$ and thus $\lim_{y \rightarrow x} f(y) = f(x)$.

\Leftarrow Now suppose the limit condition at the limit point x . Letting $\varepsilon > 0$ be given, there exists $\delta > 0$ such that if $0 < |y - x| < \delta$, then $|f(y) - f(x)| < \varepsilon$. The other case is that $y = x$ in which case $|f(y) - f(x)| = 0 < \varepsilon$. Thus f is continuous at $x \in D(f)$. ■

The problem with trying to take a limit at a point which is not a limit point of $D(f)$ is that it does not make sense. Go over the proof of why the limit is well defined and you will see this. If you are sufficiently close to a point which is not a limit point, then there will be no other points of $D(f)$ this close. Hence you could reason that any number is the limit. The concept is completely useless.

Example 4.12.8 Find $\lim_{x \rightarrow 3} \frac{x^2 - 9}{x - 3}$.

Let $x_n \rightarrow 3$, the x_n distinct, none equal to 3. Then $\frac{x_n^2-9}{x_n-3} = (x_n+3) \rightarrow 6$.

The habit students acquire of plugging in the point to take the limit is only good on useless and uninteresting limits which are not good for anything other than to give a busy work exercise.

Example 4.12.9 Let $f(x) = \frac{x^2-9}{x-3}$ if $x \neq 3$. How should f be defined at $x = 3$ so that the resulting function will be continuous there?

The limit of this function equals 6. For $x \neq 3$, $\frac{x^2-9}{x-3} = \frac{(x-3)(x+3)}{x-3} = x+3$. Therefore, by Theorem 4.12.7 it is necessary to define $f(3) \equiv 6$.

Example 4.12.10 Find $\lim_{x \rightarrow \infty} \frac{x}{1+x}$.

Write $\frac{x}{1+x} = \frac{1}{1+(1/x)}$. Now it seems clear that $\lim_{x \rightarrow \infty} 1 + (1/x) = 1 \neq 0$.

Example 4.12.11 Show $\lim_{x \rightarrow a} \sqrt{x} = \sqrt{a}$ whenever $a \geq 0$. In the case that $a = 0$, take the limit from the right.

There are two cases. First consider the case when $a > 0$. Let $\varepsilon > 0$ be given. Let $x_n \rightarrow x$ with none of the $x_n = a$. Multiply and divide by $\sqrt{x} + \sqrt{a}$. This yields

$$|\sqrt{x_n} - \sqrt{a}| = \left| \frac{x_n - a}{\sqrt{x_n} + \sqrt{a}} \right|.$$

For large n , $x_n > 0$ and so $|\sqrt{x_n} - \sqrt{a}| < \left| \frac{x_n - a}{\sqrt{a}} \right|$ which clearly converges to 0. In case $a = 0$, let $x_n \rightarrow 0$. If $\varepsilon > 0$ is given, eventually $0 < x_n < \varepsilon^2$ and so $\sqrt{x_n} < \varepsilon$ which is what it means to have $\lim_{n \rightarrow \infty} \sqrt{x_n} = 0$.

Here is a useful proposition.

Proposition 4.12.12 Suppose f is increasing on $(0, \infty)$ and is bounded above. Then $\lim_{x \rightarrow \infty} f(x) = m$ where $m \equiv \sup \{f(x) : x > 0\}$. Similar conclusions hold if ∞ is replaced with any other number. Also, if f is decreasing and bounded above, then $\lim_{x \rightarrow 0^+} f(x) = m \equiv \sup \{f(x) : x > 0\}$.

Proof: By definition, $m < \infty$ and there exists x_ε such that $m - \varepsilon < f(x_\varepsilon) \leq m$. Since f is increasing, it follows that for $y \geq x_\varepsilon$, $f(y) \in (m - \varepsilon, m]$ so $|f(y) - m| < \varepsilon$. The other claim is similar. ■

4.13 Exercises

1. Find the following limits if possible

(a) $\lim_{x \rightarrow 0^+} \frac{|x|}{x}$

(b) $\lim_{x \rightarrow 0^+} \frac{x}{|x|}$

(c) $\lim_{x \rightarrow 0^-} \frac{|x|}{x}$

(d) $\lim_{x \rightarrow 4} \frac{x^2-16}{x-4}$

(e) $\lim_{x \rightarrow 3} \frac{x^2-9}{x-3}$

(f) $\lim_{x \rightarrow -2} \frac{x^2-4}{x+2}$

(g) $\lim_{x \rightarrow \infty} \frac{x}{1+x^2}$

(h) $\lim_{x \rightarrow \infty} -2 \frac{x}{1+x^2}$

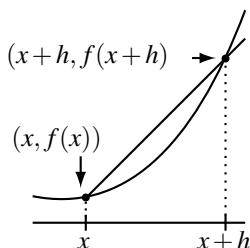
2. Find $\lim_{h \rightarrow 0} \frac{\frac{1}{(x+h)^3} - \frac{1}{x^3}}{h}$.
3. Find $\lim_{x \rightarrow 4} \frac{\sqrt[4]{x} - \sqrt{2}}{\sqrt{x} - 2}$.
4. Find $\lim_{x \rightarrow \infty} \frac{\sqrt[5]{3x} + \sqrt[4]{x} + 7\sqrt{x}}{\sqrt{3x+1}}$.
5. Find $\lim_{x \rightarrow \infty} \frac{(x-3)^{20}(2x+1)^{30}}{(2x^2+7)^{25}}$.
6. Find $\lim_{x \rightarrow 2} \frac{x^2-4}{x^3+3x^2-9x-2}$.
7. Find $\lim_{x \rightarrow \infty} \left(\sqrt{1-7x+x^2} - \sqrt{1+7x+x^2} \right)$.
8. Prove Theorem 4.12.2 for right, left and limits as $y \rightarrow \infty$.
9. Prove from the definition that $\lim_{x \rightarrow a} \sqrt[3]{x} = \sqrt[3]{a}$ for all $a \in \mathbb{R}$. **Hint:** You might want to use the formula for the difference of two cubes, $a^3 - b^3 = (a-b)(a^2 + ab + b^2)$.
10. Is it reasonable to define continuity at isolated points, those points which are not limit points, in terms of a limit?
11. Prove Theorem 4.12.7 from the definitions of limit and continuity.
12. Find $\lim_{h \rightarrow 0} \frac{(x+h)^3 - x^3}{h}$.
13. Find $\lim_{h \rightarrow 0} \frac{\frac{1}{x+h} - \frac{1}{x}}{h}$.
14. Find $\lim_{x \rightarrow -3} \frac{x^3+27}{x+3}$.
15. Find $\lim_{h \rightarrow 0} \frac{\sqrt{(3+h)^2-3}-3}{h}$ if it exists.
16. Find the values of x for which $\lim_{h \rightarrow 0} \frac{\sqrt{(x+h)^2-x}}{h}$ exists and find the limit.
17. Find $\lim_{h \rightarrow 0} \frac{\sqrt[3]{(x+h)} - \sqrt[3]{x}}{h}$ if it exists. Here $x \neq 0$.
18. Suppose $\lim_{y \rightarrow x+} f(y) = L_1 \neq L_2 = \lim_{y \rightarrow x-} f(y)$. Show $\lim_{y \rightarrow x} f(x)$ does not exist. **Hint:** Roughly, the argument goes as follows: For $|y_1 - x|$ small and $y_1 > x$, $|f(y_1) - L_1|$ is small. Also, for $|y_2 - x|$ small and $y_2 < x$, $|f(y_2) - L_2|$ is small. However, if a limit existed, then $f(y_2)$ and $f(y_1)$ would both need to be close to some number and so both L_1 and L_2 would need to be close to some number. However, this is impossible because they are different.
19. Let $f(x, y) = \frac{x^2-y^2}{x^2+y^2}$. Find $\lim_{x \rightarrow 0} (\lim_{y \rightarrow 0} f(x, y))$, $\lim_{y \rightarrow 0} (\lim_{x \rightarrow 0} f(x, y))$. If you did it right you got -1 for one answer and 1 for the other. What does this tell you about interchanging limits?
20. If f is an increasing function which is bounded above by a constant M , show that $\lim_{x \rightarrow \infty} f(x)$ exists. Give a similar theorem for decreasing functions.
21. Suppose $\{f_n\}$ is a sequence of increasing nonnegative functions defined on $[0, 1]$. Suppose also for each $x \in [0, 1]$, $\lim_{n \rightarrow \infty} f_n(x) = 0$ so you have pointwise convergence. Will it follow that f_n also converges uniformly to 0? Note that in the example where $f_n(x) = x^n$, $f_n(1)$ fails to converge to 0.

22. Show that if $\lim_{h \rightarrow 0+} \frac{f(x+h) - f(x)}{h} = m$ and $\lim_{h \rightarrow 0+} \frac{f(x) - f(x-h)}{h} = m$, then

$$\lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} = m.$$

When this happens, we say $f'(x) = m$. It is called the derivative.

23. This problem depends on the earlier problems involving the Darboux integral presented earlier, Problems 25 - 30 beginning on Page 100. The derivative at a point x , denoted as $f'(x)$ gives the slope of a tangent line to the graph of $y = f(x)$ at the point $(x, f(x))$. It is defined in terms of a limit suggested by the following picture.



This illustrates the derivative from the right which is $\lim_{h \rightarrow 0+} \frac{f(x+h) - f(x)}{h}$. The derivative from the left is defined similarly as $\lim_{h \rightarrow 0+} \frac{f(x) - f(x-h)}{h}$. Geometrically these limits give what should be defined as the slope of the tangent line to the graph of the function $y = f(x)$.

Suppose f is integrable on $[a, b]$ and x is an interior point of $[a, b]$. Suppose also that f is continuous. Suppose for all $|h|$ small enough, $\int_a^x f dt$ and $\int_a^{x+h} f dt$ both make sense. Show that

$$\lim_{h \rightarrow 0} \frac{\int_a^{x+h} f dt - \int_a^x f dt}{h} = f(x)$$

Hint: From Problems 25 - 30, the left side equals $\frac{1}{h} \int_x^{x+h} f dt$. First suppose that $h > 0$. Let M be the maximum of f on $[x, x+h]$ and let m be the minimum. Then explain why $m = \frac{1}{h} \int_x^{x+h} m dt \leq \frac{1}{h} \int_x^{x+h} f dt \leq \frac{1}{h} \int_x^{x+h} M dt = M$. By intermediate value theorem due to Bolzano, there is $y_h \in [x, x+h]$ such that $f(y_h) = \frac{1}{h} \int_x^{x+h} f dt$. Now use continuity. For the derivative from the left, apply the same argument $\frac{1}{h} \int_{x-h}^x f(t) dt$.

4.14 Videos

[continuous functions](#) [properties of continuous functions](#)

[uniform and semicontinuity](#) [limits and derivatives](#)

[approximation with polynomials](#)

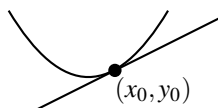
Chapter 5

The Derivative

Some functions have derivatives and some don't. Some have derivatives at some points and not at others. This chapter is on the derivative. Functions which have derivatives are better than those which don't.

5.1 The Definition of the Derivative

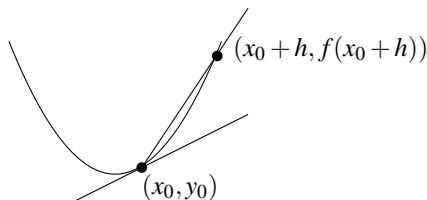
Here is a picture of the graph of a function $y = f(x)$ and a line tangent to the graph at the point (x_0, y_0) .



Thus $y_0 = f(x_0)$. Suppose m is the slope of this line. Then from algebra, the equation of the line is

$$y = y_0 + m(x - x_0) = f(x_0) + m(x - x_0)$$

The problem is to determine what m should be so that the above picture is in some sense correct. The following picture suggests how we should define m .



It seems that the slope of the line joining $(x_0 + h, f(x_0 + h))$ and $(x_0, f(x_0))$ would be getting close to m if h is small enough. Just imagine what happens as you take h smaller in this picture. This illustrates a derivative from the right because $h > 0$ in this picture. A similar picture could be drawn for negative h . This motivates the following definition of the derivative.

Definition 5.1.1 *The derivative, denoted as $f'(x_0)$, is the slope of the line tangent*

to the graph of the function $y = f(x)$ at the point $(x_0, f(x_0))$. This is defined precisely as

$$\lim_{h \rightarrow 0} \frac{f(x_0 + h) - f(x_0)}{h} \equiv f'(x_0)$$

whenever this limit exists. If you only allow positive h in the definition, then it is a derivative from the right or right derivative. If you only allow negative h in the definition, then it is a derivative from the left or left derivative. Letting $h = x - x_0$, one can also write this limit in the form $\lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0}$ where the right derivatives involve $x > x_0$ and left derivatives involve $x < x_0$. Note that

$$\lim_{h \rightarrow 0^-} \frac{f(x_0 + h) - f(x_0)}{h} = \lim_{k \rightarrow 0^+} \frac{f(x_0 - k) - f(x_0)}{-k} = \lim_{k \rightarrow 0^+} \frac{f(x_0) - f(x_0 - k)}{k}$$

Actually, it is better to express this a little differently. From the above definition, $f'(x_0)$ exists if and only if

$$\lim_{h \rightarrow 0} \frac{|f(x_0 + h) - (f(x_0) + f'(x_0)h)|}{|h|} = 0$$

if and only if $f(x_0 + h) - (f(x_0) + f'(x_0)h) = o(h)$ or $f(x_0 + h) = f(x_0) + f'(x_0)h + o(h)$ where $o(h)$ is descriptive of a function $g(h)$ with the property that $\lim_{h \rightarrow 0} \frac{g(h)}{h} = 0$. Thus we say a function $g(h)$ is $o(h)$ (little o of h) if $\lim_{h \rightarrow 0} \frac{g(h)}{h} = 0$. We use $o(h)$ as an adjective describing the behavior of a function, not as a precise description of a function. Note that, understood this way,

$$o(h) = 32o(h), o(h) - o(h) = o(h), |o(h)| = o(h), \text{ etc.}$$

This leads to the definition which I will use in what follows.

Definition 5.1.2 Let f be defined on an interval $[a, b]$. Then it is differentiable at $x \in (a, b)$ if and only if there is a constant L such that

$$f(x + h) = f(x) + Lh + o(h) \quad (5.1)$$

If h is constrained to be positive, then L is a right derivative. If h is constrained to be negative, then L is a left derivative. Then

$$L \equiv \frac{df}{dx}(x) \equiv f'(x) \equiv D_x f(x) \equiv Df(x) \equiv \dot{f}(x)$$

and we refer to this L as the derivative. Letting $x_1 + h = x_2$, an equivalent statement that $f'(x_1)$ exists is that

$$f(x_2) = f(x_1) + L(x_2 - x_1) + o(x_2 - x_1)$$

As shown above, the derivative can be considered as the slope of a tangent line, assuming such a tangent line exists. Also, at the end points, L must be a one sided derivative.

Proposition 5.1.3 There is at most one L in 5.1 so $f'(x)$ is well defined if it exists.

Proof: Suppose you have two, L, \hat{L} . Then from 5.1, $(L - \hat{L})h = o(h) - \hat{o}(h) = o(h)$. Hence $L - \hat{L} = \frac{o(h)}{h}$. Letting $h \rightarrow 0$ or $h \rightarrow 0$ from the right or the left in the case of end points, it follows $L = \hat{L}$. ■

Why bother with this little o notation? It is because this is what generalizes to higher dimensions and the notion of slope does not. It may be best to get used to it in the simpler setting of functions of one variable. As explained above, $f'(x)$ does have the geometric interpretation of being the slope of a tangent line at the point $(x, f(x))$ in one dimension.

Example 5.1.4 Let $f(x) = x^n$ where n is a nonnegative integer. Find $f'(x)$.

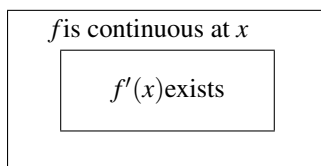
$$(x+h)^n = x^n + nx^{n-1}h + \sum_{k=2}^n \binom{n}{k} x^{n-k}h^k = x^n + nx^{n-1} + o(h)$$

Thus $f'(x) = nx^{n-1}$.

Example 5.1.5 Let $f(x) = |x|$. Show $f'(0)$ does not exist.

If it did exist, then $|h| = Lh + o(h)$ for some L . However, taking $h > 0$ and letting $h \rightarrow 0$ yields $L = 1$ and letting $h < 0$ and $h \rightarrow 0$ yields $L = -1$. Note that this function does have a right and a left derivative at 0.

The following diagram shows how continuity at a point and differentiability there are related.



Theorem 5.1.6 If f is defined near x and f is differentiable at x then f is continuous at x . Also if $f'(x)$ exists for some x in $[a, b]$, then

$$o(f(x+h) - f(x)) = o(h)$$

Proof: Suppose $\lim_{n \rightarrow \infty} x_n = x$. Does it follow that $\lim_{n \rightarrow \infty} f(x_n) = f(x)$? By assumption,

$$f(x) - f(x_n) = f'(x)(x - x_n) + o(x - x_n)$$

Now from the definition of $o(x - x_n)$, $|o(x - x_n)| < |x - x_n|$ if n is large enough. Hence, for large n , $|f(x) - f(x_n)| \leq (|f'(x)| + 1)|x_n - x|$ and so by the squeezing theorem,

$$\lim_{n \rightarrow \infty} f(x_n) = f(x).$$

Then by Theorem 4.0.8, f is continuous at x .

Consider the other claim. Let $\varepsilon > 0$ be given. Let

$$H(h) \equiv \begin{cases} \frac{o(f(x+h) - f(x))}{f(x+h) - f(x)} & \text{if } f(x+h) - f(x) \neq 0 \\ 0 & \text{if } f(x+h) - f(x) = 0 \end{cases}$$

$$\left| \frac{o(f(x+h) - f(x))}{h} \right| = |H(h)| \left| \frac{f(x+h) - f(x)}{h} \right|$$

Now since $f'(x)$ exists, $|f(x+h) - f(x)| \leq |f'(x)| |h| + |h| \leq (1 + |f'(x)|) |h| \equiv C|h|$ for all h small enough. Hence, for small $|h|$,

$$\left| \frac{f(x+h) - f(x)}{h} \right| \leq C, \quad \left| \frac{o(f(x+h) - f(x))}{h} \right| \leq C|H(h)|$$

Now $\lim_{h \rightarrow 0} H(h) = 0$ and so $o(f(x+h) - f(x)) = o(h)$. For derivatives from the right or left, you simply constrain h to be either positive or negative and there is no change. ■

Weierstrass gave an example of a function continuous at every point yet differentiable at no point, but you can easily see the example of $y = |x|$ which is continuous at 0 but not differentiable there. To see a standard example of a nowhere differentiable continuous function, see my single variable advanced calculus book.

5.2 Finding the Derivative

Obviously there need to be simple ways of finding the derivative when it exists. There are rules of derivatives which make finding the derivative very easy. In the following theorem, the derivative could refer to right or left derivatives as well as regular derivatives.

Theorem 5.2.1 *Let a, b be numbers and suppose $f'(t)$ and $g'(t)$ exist. Then the following formulas are obtained.*

$$(af + bg)'(t) = af'(t) + bg'(t). \quad (5.2)$$

$$(fg)'(t) = f'(t)g(t) + f(t)g'(t). \quad (5.3)$$

The formula, 5.3 is referred to as the product rule.

If $f'(g(t))$ exists and $g'(t)$ exists, then $(f \circ g)'(t)$ also exists and

$$(f \circ g)'(t) = f'(g(t))g'(t).$$

This is called the chain rule. In this rule, for the sake of simplicity, assume the derivatives are real derivatives, not derivatives from the right or the left. If $f(t) = t^n$ where n is any integer, then

$$f'(t) = nt^{n-1}. \quad (5.4)$$

Also, whenever $f'(t)$ exists, $f'(t) = \lim_{h \rightarrow 0} \frac{f(t+h) - f(t)}{h}$ where this definition can be adjusted in the case where the derivative is a right or left derivative by letting $h > 0$ or $h < 0$ only and considering a one sided limit. This is equivalent to $f'(t) = \lim_{s \rightarrow t} \frac{f(s) - f(t)}{t - s}$ with the limit being one sided in the case of a left or right derivative.

Proof: 5.2 is left for you. Consider 5.3

$$\begin{aligned} fg(t+h) - fg(t) &= (f(t) + f'(t)h + o(h))(g(t) + g'(t)h + o(h)) - f(t)g(t) \\ &= f'(t)g(t)h + g'(t)f(t)h + o(h) \end{aligned}$$

This shows 5.3.

Next consider the chain rule. By Theorem 5.1.6

$$f(g(t+h)) - f(g(t)) = f'(g(t))(g(t+h) - g(t)) + o(g(t+h) - g(t))$$

$$= f'(g(t))(g'(t)h + o(h)) + o(h) = f'(g(t))g'(t)h + o(h)$$

The last claim follows from Example 5.1.4 in case n is a positive integer. If n is 0, then the claim is obvious because the function is a constant so its derivative is 0. It remains to consider the case where n is a negative integer. First consider $f(t) = t^{-1}$. Then

$$\frac{f(t+h) - f(t)}{h} = \frac{\frac{1}{t+h} - \frac{1}{t}}{h} = -\frac{1}{(t+h)t}$$

For all $t \neq 0$, the limit of this last expression is $-\frac{1}{t^2}$ using the properties of the limit. Therefore, if $f(t) = t^{-n}$, then $f(t) = (t^n)^{-1}$ and so, by the chain rule and what was just shown for positive exponent n , $f'(t) = (-1)(t^n)^{-2}nt^{n-1} = -nt^{-(n+1)}$ showing that the claim holds in this case also. ■

Corollary 5.2.2 *Let $f'(t), g'(t)$ both exist and $g(t) \neq 0$, then the quotient rule holds.*

$$\left(\frac{f}{g}\right)' = \frac{f'(t)g(t) - f(t)g'(t)}{g(t)^2}$$

Proof: This is left to you. Use the chain rule and the product rule. ■

Higher order derivatives are defined in the obvious way. $f'' \equiv (f')'$ etc. Also the Leibniz notation is defined by $\frac{dy}{dx} = f'(x)$ where $y = f(x)$ and the second derivative is denoted as $\frac{d^2y}{dx^2}$ with various other higher order derivatives defined similarly. When people write $y^{(n)}$ they mean the n^{th} derivative. Similarly $f^{(n)}(x)$ refers to the n^{th} derivative.

The chain rule has a particularly attractive form in Leibniz's notation. Suppose $y = g(u)$ and $u = f(x)$. Thus $y = g \circ f(x)$. Then from the above theorem

$$(g \circ f)'(x) = g'(f(x))f'(x) = g'(u)f'(x)$$

or in other words, $\frac{dy}{dx} = \frac{dy}{du} \frac{du}{dx}$. Notice how the du cancels. This particular form is a very useful crutch and is used extensively in applications.

5.3 Derivatives of Inverse Functions

It happens that if f is a differentiable one to one function defined on an interval, $[a, b]$, and $f'(x)$ exists and is non zero then the inverse function f^{-1} has a derivative or one sided derivative at the point $f(x)$.

Theorem 5.3.1 *Let $f : [a, b] \rightarrow \mathbb{R}$ be continuous and one to one. Suppose $f'(x)$ exists for some $x \in [a, b]$ and $f'(x) \neq 0$, a one sided derivative at the end points. Then $(f^{-1})'(f(x))$ exists and is given by the formula, $(f^{-1})'(f(x)) = \frac{1}{f'(x)}$.*

Proof: By Lemma 4.4.3, and Corollary 4.5.1 on Page 112 f is either strictly increasing or strictly decreasing and f^{-1} is continuous on an interval $f([a, b])$. Constrain h to have the appropriate sign if at an endpoint of $f([a, b])$, and letting $|h|$ be sufficiently small otherwise, let x be a point where $f'(x) \neq 0$ and $f(x) = y$

$$h = f(f^{-1}(y+h)) - f(f^{-1}(y)) =$$

$$f'(x)(f^{-1}(y+h) - f^{-1}(y)) + o(f^{-1}(y+h) - f^{-1}(y)) \quad (*)$$

By continuity of f^{-1} , $|o(f^{-1}(y+h) - f^{-1}(y))| < \frac{1}{2}|f'(x)||f^{-1}(y+h) - f^{-1}(y)|$ if h is small enough and so, from the triangle inequality in $*$,

$$|h| \geq \frac{1}{2}|f'(x)||f^{-1}(y+h) - f^{-1}(y)|,$$

$$\frac{|o(f^{-1}(y+h) - f^{-1}(y))|}{|h|} \leq \frac{2|o(f^{-1}(y+h) - f^{-1}(y))|}{|f'(x)||f^{-1}(y+h) - f^{-1}(y)|}$$

showing that $o(f^{-1}(y+h) - f^{-1}(y)) = o(h)$. From $*$,

$$\frac{1}{f'(x)}h + o(h) = f^{-1}(y+h) - f^{-1}(y) = f^{-1}(f(x)+h) - f^{-1}(f(x))$$

Which proves the theorem. ■

This is one of those theorems which is very easy to remember if you neglect the difficult questions and simply focus on formal manipulations. Consider the following. $f^{-1}(f(x)) = x$. Now use the chain rule to write $(f^{-1})'(f(x))f'(x) = 1$, and then divide both sides by $f'(x)$ to obtain $(f^{-1})'(f(x)) = \frac{1}{f'(x)}$. Of course this gives the conclusion of the above theorem rather effortlessly and it is formal manipulations like this which aid in remembering formulas such as the one given in the theorem.

Example 5.3.2 Let $f(x) = 8 + x^2 + x^3 + 7x$. Show that f has an inverse and find $(f^{-1})'(8)$.

I am not able to find a formula for the inverse function. This is typical in useful applications so you need to get used to this idea. The methods of algebra are insufficient to solve hard problems in analysis. You need something more. The question is to determine whether f has an inverse. To do this, $f'(x) = 2x + 3x^2 + 7 > 0$ for all x . By Corollary 5.11.5 on Page 148, this function is strictly increasing on \mathbb{R} and so it has an inverse function although I have no idea how to find an explicit formula for this inverse function. However, I can see that $f(0) = 8$ and so by the formula for the derivative of an inverse function,

$$(f^{-1})'(8) = (f^{-1})'(f(0)) = \frac{1}{f'(0)} = \frac{1}{7}.$$

In practice, we typically don't bother with the mathematical details. We have $f(x) = y$ and the inverse function is of the form $x = f^{-1}(y)$. Thus it involves finding $\frac{dx}{dy}(y) \equiv (f^{-1})'(y)$. The existence of the derivative of the inverse function exists by the above argument. Therefore, all that remains is to use the chain rule. Take $\frac{d}{dy}$ of both sides of $f(x) = y$, $f'(x)\frac{dx}{dy} = 1$. Thus

$$\frac{dx}{dy}(y) \equiv (f^{-1})'(y) = \frac{1}{f'(x)} = \frac{1}{f'(f^{-1}(y))} = \frac{1}{f'(x)}$$

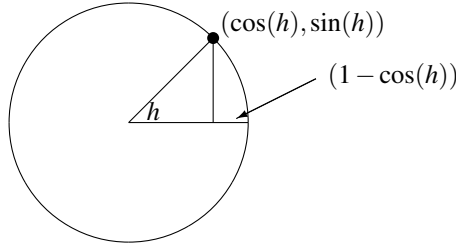
which is the same as obtained earlier. You know the inverse has a derivative and so it suffices to use the chain rule.

5.4 Circular Functions and Inverses

Here in this section, the derivatives of the circular functions are derived and then the derivatives of their inverse functions are considered.

Theorem 5.4.1 $\sin'(x) = \cos(x), \cos'(x) = -\sin(x)$.

Proof: Consider the picture where here h is small



From Corollary 2.3.9

$$(1 - \cos(h)) + \sin(h) \geq h \geq \sin(h)$$

It follows that

$$\frac{\sin(h)}{1 + \cos(h)} = \frac{1 - \cos^2(h)}{\sin(h)(1 + \cos(h))} = \frac{1 - \cos(h)}{\sin(h)} + 1 \geq \frac{h}{\sin(h)} \geq 1 \quad (5.5)$$

and so $\lim_{h \rightarrow 0} \frac{h}{\sin(h)} = \lim_{h \rightarrow 0} \frac{\sin(h)}{h} = 1$.

$$\frac{1 - \cos(h)}{h} = \frac{\sin^2(h)}{h(1 + \cos(h))} = \frac{\sin(h)}{h} \frac{\sin(h)}{1 + \cos(h)} \rightarrow 0$$

$$\frac{h - \sin(h)}{h} = \frac{h}{h} - \frac{\sin(h)}{h} \rightarrow 0$$

so $1 - \cos(h) = o(h)$ and $h - \sin(h) = o(h)$. Then

$$\begin{aligned} \sin(x+h) - \sin(x) &= \sin(x)\cos(h) + \cos(x)\sin(h) - \sin(x) \\ &= \sin(x)(\cos(h) - 1) + \cos(x)(\sin(h) - h) + \cos(x)h \\ &= \cos(x)h + o(h) \end{aligned}$$

so $\sin'(x) = \cos(x)$.

$$\cos(x+h) - \cos(x) = \cos(x)(\cos(h) - 1) - \sin(x)\sin(h)$$

$$= o(h) - \sin(x)(\sin(h) - h) - \sin(x)h = -\sin(x)h + o(h)$$

Thus $\cos'(x) = -\sin(x)$. ■

The sine function is one to one on $[-\frac{\pi}{2}, \frac{\pi}{2}]$ taking all values between -1 and 1 and so one can define

$$\arcsin : [-1, 1] \rightarrow \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$$

as an inverse function for the sine restricted to $[-\frac{\pi}{2}, \frac{\pi}{2}]$. In words, $\arcsin(y)$ is the angle whose sine is y which lies in $[-\frac{\pi}{2}, \frac{\pi}{2}]$. Letting $\sin(x) = y$, you can find $\frac{dx}{dy}$ by using the

chain rule. Thus $\cos(x) \frac{dx}{dy} = 1$ and so $\frac{dx}{dy} = \frac{1}{\cos(x)}$. Now for $x \in [-\frac{\pi}{2}, \frac{\pi}{2}]$, $\cos(x) \geq 0$ and so the above equation reduces to

$$\arcsin'(y) \equiv \frac{dx}{dy} = \frac{1}{\sqrt{1 - \sin^2(x)}} = \frac{1}{\sqrt{1 - y^2}}$$

The cosine function is one to one on $[0, \pi]$ taking all values between -1 and 1 and so one can define

$$\arccos : [-1, 1] \rightarrow [0, \pi]$$

as an inverse function for the cosine restricted to $[0, \pi]$. In words, $\arccos(y)$ is the angle whose cosine is y which lies in $[0, \pi]$. Letting $y = \cos(x)$, you can find $\frac{dx}{dy}$ using the chain rule. As explained, this is the derivative of the inverse function just described. $1 = -\sin(x) \frac{dx}{dy}$ and so $\frac{dx}{dy} = -\frac{1}{\sin(x)}$. For $x \in [0, \pi]$, $\sin(x) \geq 0$ and so

$$-\sin(x) = -\sqrt{1 - \cos^2(x)} = -\sqrt{1 - y^2}$$

Thus

$$\arccos'(y) \equiv \frac{dx}{dy} = -\frac{1}{\sqrt{1 - \cos^2(x)}} = -\frac{1}{\sqrt{1 - y^2}}$$

The tangent function is one to one on $(-\frac{\pi}{2}, \frac{\pi}{2})$ and maps onto $(-\infty, \infty)$, all of \mathbb{R} . Thus one can define $\arctan(y)$ as $x \in (-\frac{\pi}{2}, \frac{\pi}{2})$ where $\tan(x) = y$ as the above. Now applying the quotient rule to find $\tan'(x)$,

$$\tan'(x) = \frac{\cos^2(x) - (-\sin(x))\sin(x)}{\cos^2(x)} = \frac{1}{\cos^2(x)} = \sec^2(x) = 1 + \tan^2(x)$$

the last being a well known identity which says essentially that $\cos^2(x) + \sin^2(x) = 1$. Then as before, $y = \tan(x)$,

$$1 = (1 + \tan^2(x)) \frac{dx}{dy} = (1 + y^2) \frac{dx}{dy}$$

and so $\arctan'(y) = \frac{1}{1+y^2}$. You can do all the other trigonometric functions and their inverses the same way. Of course none of them have inverses unless their domains are restricted as above. For example, $x \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ in order that \sin will be one to one. The choice of interval on which the function is one to one is somewhat arbitrary. One could have done the same thing for \arcsin if the interval had been $[\frac{3\pi}{2}, \frac{5\pi}{2}]$ instead of $[-\frac{\pi}{2}, \frac{\pi}{2}]$ for example. However, it is traditional to pick the interval to which the function is restricted to be that interval closest to 0 such that the function is one to one and maps onto its maximum range. This is done to maximize the usefulness of the definition. The following table summarizes the derivatives of the trigonometric functions and their inverses. In the table D will be the domain of the function and R will be the range of the function.

Table of Derivatives

$f(x)$	Domain	Range	$f'(x)$
$\sin(x)$	\mathbb{R}	$[-1, 1]$	$\cos(x)$
$\arcsin(x)$	$[-1, 1]$	$[-\frac{\pi}{2}, \frac{\pi}{2}]$	$\frac{1}{\sqrt{1-x^2}}$
$\cos(x)$	\mathbb{R}	$[-1, 1]$	$-\sin(x)$
$\tan(x)$	all except odd multiples of $\frac{\pi}{2}$	\mathbb{R}	$\sec^2(x)$
$\arctan(x)$	\mathbb{R}	$(-\frac{\pi}{2}, \frac{\pi}{2})$	$\frac{1}{1+x^2}$
$\sec(x)$	all except odd multiples of $\frac{\pi}{2}$	$[1, \infty) \cup (-\infty, -1]$	$\sec(x) \tan(x)$
$\operatorname{arcsec}(x)$	$[1, \infty) \cup (-\infty, -1]$	$[0, \frac{\pi}{2}) \cup (\frac{\pi}{2}, \pi]$	$\frac{1}{ y \sqrt{y^2-1}}$

For the line corresponding to arcsec , I have picked the domain to be $[1, \infty) \cup (-\infty, -1]$ to correspond to restricting $\sec(x)$ to $[0, \frac{\pi}{2}) \cup (\frac{\pi}{2}, \pi]$. There is no consensus on how to do this. I have done it this way because the interval on which $\cos(x)$ is one to one which we use in defining the inverse cosine is $[0, \pi]$ and I have tried to make it as similar to this as possible.

One can do similar things for \csc , and arccsc and \cot and arccot but the above is likely enough of this tedium to cover the situations of interest.

5.5 Exponential Functions and Logarithms

First consider $\ln(x)$. Recall how, for $x > 1$ it is the area between 1, x , which lies under the graph of the curve $y = 1/t$. Thus for $x > 1$ and small positive h ,

$$\ln(x+h) - \ln(x)$$

is the area between x and $x+h$ under the graph of the function $y = 1/t$. Therefore,

$$h \frac{1}{x+h} \leq \ln(x+h) - \ln(x) \leq h \left(\frac{1}{x} \right)$$

Divide by h to get

$$\frac{1}{x+h} \leq \frac{\ln(x+h) - \ln(x)}{h} \leq \frac{1}{x}.$$

Then by the squeezing theorem, $\lim_{h \rightarrow 0+} \frac{\ln(x+h) - \ln(x)}{h} = \frac{1}{x}$. If $h < 0$, a similar argument shows

$$\lim_{h \rightarrow 0-} \frac{\ln(x+h) - \ln(x)}{h} = \lim_{h \rightarrow 0+} \frac{\ln(x) - \ln(x-h)}{h} = \frac{1}{x}$$

See Problem 22 on Page 127.

Actually one should also consider $\ln(|x|)$ because this allows the consideration of negative values of x . For $x < 0$, $\ln(|x|) = \ln(-x)$ and so, by the chain rule, the derivative of this function of x is $\frac{1}{-x}(-1) = \frac{1}{x}$.

Thus \ln is differentiable as is $x \rightarrow \ln|x|$ for all $x \neq 0$. It follows from Theorem 5.3.1 that the inverse function \exp is also differentiable. Now by definition, $\ln(\exp(x)) = x$ and so, by the chain rule,

$$\frac{1}{\exp(x)} \exp'(x) = 1$$

and so $\exp'(x) = \exp(x)$. The fact that \exp' is equal to \exp turns out to be very significant in finding solutions to differential equations.

Suppose $b > 0, b \neq 1$, and $f(x) = b^x$. What is $f'(x)$? You know that $b^x = \exp(x \ln(b))$ and so by the chain rule,

$$f'(x) = \exp(x \ln(b)) \ln(b) = \ln(b) b^x$$

If $f(x) = \log_b(x)$, what is $f'(x)$? $f(x) \equiv \frac{\ln(x)}{\ln(b)}$ and so $f'(x) = \frac{1}{\ln(b)} \frac{1}{x}$. Also we can consider $f(x) = x^r$ for $x > 0$ and r a fixed positive real number. Then $f(x) \equiv \exp(r \ln(x))$ and so, by the chain rule,

$$f'(x) = \exp(r \ln(x)) \frac{r}{x} \equiv r x^{r-1}$$

Table of Derivatives

$f(x)$	Domain	Range	$f'(x)$
$\ln(x)$	$(0, \infty)$	\mathbb{R}	$\frac{1}{x}$
$\log_b(x), b \neq 1$	$(0, \infty)$	\mathbb{R}	$\frac{1}{\ln(b)} \frac{1}{x}$
$\exp(x)$	\mathbb{R}	$(0, \infty)$	$\exp(x)$
e^x	\mathbb{R}	$(0, \infty)$	e^x
$b^x, b \neq 1, b > 0$	\mathbb{R}	$(0, \infty)$	$\ln(b) b^x$
$x^r, r > 0$	$(0, \infty)$	$(0, \infty)$	$r x^{r-1}$
$\ln(x)$	$(0, \infty) \cup (-\infty, 0)$	\mathbb{R}	$\frac{1}{x}$
$\cosh(x)$	\mathbb{R}	$[1, \infty)$	$\sinh(x)$
$\sinh(x)$	\mathbb{R}	\mathbb{R}	$\cosh(x)$

Note that if $r < 0, x^{-r} \equiv (x^r)^{-1}$ and so, by the chain rule, the derivative is

$$(-1)(x^r)^{-2} r x^{r-1} = -r x^{-(r+1)}$$

thanks to Proposition 2.6.3 which says the usual rules of exponents hold. The last two lines are left as exercises.

5.6 The Complex Exponential

It was shown in introductory topics that every complex number can be written in the form $r(\cos \theta + i \sin \theta)$ where $r \geq 0$. See Section 1.14. Laying aside the zero complex number, this shows that every non zero complex number is of the form $e^{\alpha}(\cos \beta + i \sin \beta)$. We write

this in the form $e^{\alpha+i\beta}$. When you have a function $f(t) = u(t) + iv(t)$, $f'(t)$ is defined as $u'(t) + iv'(t)$ if and only if the derivatives of the real functions u and v exist.

Having made the above definitions, does it follow that the expression for $e^{(\alpha+i\beta)t}$ preserves the most important property of the function $t \rightarrow e^{(\alpha+i\beta)t}$ for t real, that $(e^{(\alpha+i\beta)t})' = (\alpha + i\beta)e^{(\alpha+i\beta)t}$? By the definition just given which does not contradict the usual definition in case $\beta = 0$ and the usual rules of differentiation in calculus,

$$\begin{aligned} (e^{(\alpha+i\beta)t})' &\equiv (e^{\alpha t} (\cos(\beta t) + i \sin(\beta t)))' \\ &= e^{\alpha t} [\alpha (\cos(\beta t) + i \sin(\beta t)) + (-\beta \sin(\beta t) + i\beta \cos(\beta t))] \end{aligned}$$

Now consider the other side. From the definition it equals

$$\begin{aligned} (\alpha + i\beta) (e^{\alpha t} (\cos(\beta t) + i \sin(\beta t))) &= e^{\alpha t} [(\alpha + i\beta) (\cos(\beta t) + i \sin(\beta t))] \\ &= e^{\alpha t} [\alpha (\cos(\beta t) + i \sin(\beta t)) + (-\beta \sin(\beta t) + i\beta \cos(\beta t))] \end{aligned}$$

which is the same thing. This is of fundamental importance in differential equations. It shows that there is no change in going from real to complex numbers for ω in the consideration of the problem $y' = \omega y$, $y(0) = 1$. The solution is always $e^{\omega t}$. The formula just discussed, that $e^{\alpha} (\cos \beta + i \sin \beta) = e^{\alpha+i\beta}$ is Euler's formula.

5.7 Related Rates and Implicit Differentiation

Related rates problems involve variables which are related by some expression and you know the rate at which all but one of the variables are changing. Then the idea is to find how fast the other variable is changing. The relation and rules of differentiation give a relation between their derivatives and enable you to obtain the information.

Example 5.7.1 A point moves along the curve $xy = 8$ and it is observed that at the point $(2, 4)$, $\frac{dx}{dt} = 3$. Find $\frac{dy}{dt}$ at this point.

This is a related rate problem, the relation between the variables being $xy = 8$. Thus each variable is really a function of t . By the product rule, $x'y + y'x = 0$ and at the point of interest, certain things are known. Substituting these into the above equation gives $3(4) + y'(2) = 0$. Then you solve for y' . Thus $y' = -6$.

Example 5.7.2 The volume of a ball of radius r is given by $V = \frac{4}{3}\pi r^3$. Suppose r is a function of t . It is observed that $\frac{dV}{dt} = 2\pi$ when the radius equals 4. Find $\frac{dr}{dt}$ when the radius is 4.

You have from the relation, $V' = 4\pi r^2 r'$. Now insert the given information. $2\pi = 4\pi(16)r'$. Then solve for r' to find that $r' = 1/32$.

A similar process is implicit differentiation which I will illustrate with an example.

Example 5.7.3 Suppose $y^3x^2 + 3xy + y^4 = 5$. Assuming the relation defines y as a function of x , determine $y'(x)$ at the point $(1, 1)$. Of course one should wonder whether the relation really does define y as a function of x . This is all part of the implicit function theorem in Section 24.

See how this is similar to the related rates problems in the sense that you have a relation between two variables. Differentiate with respect to x regarding y as a function of x . Then $3y^2y'x^3 + y^3(2x) + 3y + 3xy' + 4y^4y' = 0$. Here y' means $\frac{dy}{dx}$. Now solve for y' . This yields $y' = -\frac{3y+2xy^3}{3x+3x^3y^2+4y^4}$. At the point of interest, $y' = -\frac{3+2}{3+3+4} = -\frac{1}{2}$. It turns out that this formal manipulation is perfectly all right provided the denominator in the above formula is not 0 and this is what the implicit function theorem says.

5.8 Exercises

- Verify the last two lines in Table 5.5.
- In each of the following, assume the relation defines y as a function of x for values of x and y of interest and find $y'(x)$. This illustrates the technique of implicit differentiation.

(a) $xy^2 + \sin(y) = x^3 + 1$	(f) $\sqrt{x^2 + y^4} \sin(y) = 3x$
(b) $y^3 + x \cos(y^2) = x^4$	(g) $y^3 \sin(x) + y^2 x^2 = 2^{x^2} y + \ln y $
(c) $y \cos(x) = \tan(y) \cos(x^2) + 2$	(h) $y^2 \sin(y)x + \log_3(xy) = y^2 + 11$
(d) $(x^2 + y^2)^6 = x^3 y + 3$	(i) $\sin(x^2 + y^2) + \sec(xy) = e^{x+y} + y^{2y} + 2$
(e) $\frac{xy^2 + y}{y^5 + x} + \cos(y) = 7$	(j) $\sin(\tan(xy^2)) + y^3 = 16$
- In each of the following, assume the relation defines y as a function of x for values of x and y of interest. Use the chain rule to show y satisfies the given differential equation.

(a) $x^2 y + \sin y = 7, (x^2 + \cos y) y' + 2xy = 0$.	(b) $x^2 y^3 + \sin(y^2) = 5, 2xy^3 + (3x^2 y^2 + 2(\cos(y^2))y) y' = 0$.
(c) $y^2 \sin(y) + xy = 6, (2y(\sin(y)) + y^2(\cos(y)) + x) y' + y = 0$.	
- Suppose $f(x+y) = f(x) + f(y)$ and f is continuous at 0. Find all solutions to this functional equation which are continuous at $x = 0$. Now find all solutions which are bounded near 0.
- Suppose $f(x+y) = f(x)f(y)$ and f is differentiable and not identically zero. Find all solutions to this functional equation. **Hint:** First show the functional equation requires $f > 0$.
- Suppose $f(xy) = f(x) + f(y)$ for $x, y > 0$. Suppose also f is differentiable. Find all solutions to this functional equation.
- The volume of a cylinder is $\pi r^2 h$ and suppose it equals a constant value of 6π but that $\frac{dh}{dt} = 2$ when $r = 4$. Find $\frac{dr}{dt}$ when $r = 4$.
- Let $V = \pi r^2 h$ be the volume of a cylinder of radius r and height h . Suppose it is observed that $\frac{dV}{dt} = 2\pi, \frac{dr}{dt} = 2$ when $r = 2$ and $h = 4$. Determine $\frac{dh}{dt}$.

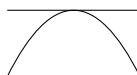
9. The ideal gas law is of the form $PV = kT$ where k is a constant depending on the gas and number of moles, T is the temperature, P the pressure, and V the volume. Assuming these variables are all functions of t , suppose it is observed at some time that $T'(t) = 5$, $V(t) = 1$, $P(t) = 2$, $P'(t) = 2$. Determine $V'(t)$ at this time.
10. Bernoulli's law states that in an incompressible fluid, $\frac{v^2}{2g} + z + \frac{P}{\gamma} = C$ where C is a constant. Here v is the speed, P is the pressure, and z is the height above some reference point. The constants g and γ are the acceleration of gravity and the weight density of the fluid. Suppose measurements indicate that $\frac{dv}{dt} = -3$, and $\frac{dz}{dt} = 2$. Find $\frac{dP}{dt}$ when $v = 7$ in terms of g and γ .

5.9 Local Extreme Points

When you are on top of a hill, you are at a local maximum although there may be other hills higher than the one on which you are standing. Similarly, when you are at the bottom of a valley, you are at a local minimum even though there may be other valleys deeper than the one you are in. The word, “local” is applied to the situation because if you confine your attention only to points close to your location, you are indeed at either the top or bottom.

Definition 5.9.1 Let $f : D(f) \rightarrow \mathbb{R}$ where here $D(f)$ is only assumed to be some subset of \mathbb{R} . Then $x \in D(f)$ is a local minimum (maximum) if there exists $\delta > 0$ such that whenever $y \in B(x, \delta) \cap D(f)$, it follows $f(y) \geq (\leq) f(x)$. The plural of minimum is minima and the plural of maximum is maxima.

Derivatives can be used to locate local maxima and local minima.



Note how the tangent line is horizontal. If you were not at a local maximum or local minimum, the function would be falling or climbing and the tangent line would not be horizontal.

Theorem 5.9.2 Suppose $f : U \rightarrow \mathbb{R}$ where U is an open subset of \mathbb{R} and suppose $x \in U$ is a local maximum or minimum and $f'(x)$ exists. Then $f'(x) = 0$.

Proof: Since U is an open set, there exists $\delta > 0$ such that $(x - \delta, x + \delta) \subseteq U$. Now if x is a local minimum,

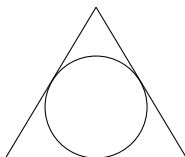
$$\begin{aligned} f'(x) &= \lim_{h \rightarrow 0^+} \frac{f(x+h) - f(x)}{h} \geq 0 \\ f'(x) &= \lim_{h \rightarrow 0^-} \frac{f(x+h) - f(x)}{h} \leq 0 \end{aligned}$$

Therefore, $f'(x) = 0$. The case where x is a local maximum is similar. You just turn around the inequality signs in the above. ■

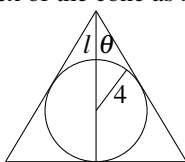
Points at which the derivative of a function equals 0 are sometimes called critical points. Included in the set of critical points are those points where f' fails to exist. You could end up with a local maximum or minimum at such a point. Think of $y = |x|$. When $x = 0$ no derivative exists and it is a local minimum.

The following is a typical minimization problem, this one heavily dependent on geometry.

Example 5.9.3 Find the volume of the smallest right circular cone which can be circumscribed about a ball of radius 4 inches. Such a cone has volume equal to $\frac{\pi}{3}r^2h$ where r is the radius of the cone and h the height.



Consider the following picture of a cross section in which l is the length of the line from the center of the ball to the vertex of the cone as shown.



The angle between the indicated radius of length 4 and the side of the cone is $\pi/2$ from geometric considerations. Thus $l \sin \theta = 4$ and the volume of the cone is

$$\frac{\pi}{3} (l+4) ((l+4) \tan(\theta))^2.$$

2θ is no more than π and so $\theta < \pi/2$. Thus $\tan \theta$ is positive and equals $\frac{\sin \theta}{\sqrt{1-\sin^2 \theta}} = \frac{4/l}{\sqrt{1-(4/l)^2}} = \frac{4}{\sqrt{l^2-16}}$. Then the volume of the cone is

$$\frac{\pi}{3} (l+4) \left((l+4) \frac{4}{\sqrt{l^2-16}} \right)^2 = \frac{16}{3} \frac{\pi}{l^2-16} (l+4)^3$$

It seems there should be a solution to this problem and so we only have to find it by taking a derivative and setting it equal to 0 because the solution will surely be a local minimum. To take the derivative, use the rules of differentiation developed above. The derivative is $-\frac{16}{3} \frac{\pi}{(l-4)^2} (-l^2 + 8l + 48)$. Obviously you cannot have $l = 4$. Such a situation would not even give a triangle. Therefore, the solution to the problem involves $l^2 - 8l - 48 = 0$. There are two solutions, $l = 12$ or $l = -4$, the latter making absolutely no sense at all. Hence $l = 12$ must be the answer and the height of the cone is 16. The minimum volume is then $\frac{16}{3} \frac{\pi}{12^2-16} (12+4)^3 = \frac{512}{3} \pi$. I think you probably could not do this problem without the methods of calculus.

Now here is an example about minimizing cost. It is another example which you could not work without the methods of calculus.

Exercise 5.9.4 A cylindrical can is to have volume 20π cubic inches. The top costs 2 cents per square inch and the sides cost 1 cent per square inch. What is the radius of the can which costs as little as possible.

You need $20\pi = \pi r^2 h$ and so $r^2 h = 20$. Now the cost is $C = 2\pi r h + 2\pi r^2 (2)$. Then the total cost in terms of r is $C = 40 \frac{\pi}{r} + 4\pi r^2$. Thus, taking the derivative and setting equal to 0 yields the radius which minimizes the cost is $\sqrt[3]{5}$ inches.

This scheme in which you take the derivative and set it equal to zero might not find the answer. It only gives candidates for the answer on the interior of an interval. Perhaps, like the above two examples these are the only points of interest. However, in general, when you look for the absolute maximum or minimum, you must consider the end points of the interval also.

Example 5.9.5 Let $f(x) = x^3 - 3x$ for $x \in [0, 4]$. Find the maximum and minimum of this function on this interval.

There exists a maximum and a minimum by the extreme value theorem. These could occur on $(0, 4)$ or at an end point. To find possibilities on $(0, 4)$, take the derivative and set equal to 0. $3x^2 - 3 = 0, x = 1$. Now $x = 1, 0, 4$ are all possibilities. $f(0) = 0, f(1) = -2, f(4) = 52$. Thus the maximum occurs at the right endpoint and is 52. The minimum occurs when $x = 1$ and is -2 .

The above illustrates how it is done in general. You consider critical points and end points. Then among these points, you find the one which gives the best answer.

5.10 Exercises

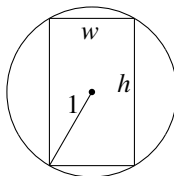
1. If $f'(x) = 0$, is it necessary that x is either a local minimum or local maximum?
Hint: Consider $f(x) = x^3$.
2. A continuous function f defined on $[a, b]$ is to be maximized. It was shown above in Theorem 5.9.2 that if the maximum value of f occurs at $x \in (a, b)$, and if f is differentiable there, then $f'(x) = 0$. However, this theorem does not say anything about the case where the maximum of f occurs at either a or b . Describe an inequality which will be satisfied at the point where f achieves its maximum on $[a, b]$ assuming f' exists on $[a, b]$. Describe an inequality which will locate the minimum on $[a, b]$ also under the assumption that f' exists. **Hint:** If f achieves its maximum at an interior point x , then $f'(x)(x - y) = 0$ for all $y \in [a, b]$. Look for something like this except with an inequality rather than an equal sign.
3. Let $y = x^x$ for $x \in (0, \infty)$. Find $y'(x)$.
4. Show using the product rule that $y'(x) = 2x$ for $y(x) = x^2$. Now use induction to verify that for n a positive integer, if $y(x) = x^n$, then $y'(x) = nx^{n-1}$.
5. Find the maximum and minimum values and the values of x where these are achieved for the function $f(x) = x + \sqrt{25 - x^2}$.
6. A piece of wire of length L is to be cut in two pieces. One piece is bent into the shape of an equilateral triangle and the other piece is bent to form a square. How should the wire be cut to maximize the sum of the areas of the two shapes? How should the wire be bent to minimize the sum of the areas of the two shapes? **Hint:** Be sure to consider the case where all the wire is devoted to one of the shapes separately. This is a possible solution even though the derivative is not zero there.
7. Lets find the point on the graph of $y = \frac{x^2}{4}$ which is closest to $(0, 1)$. One way to do it is to observe that a typical point on the graph is of the form $(x, \frac{x^2}{4})$ and then

to minimize the function $f(x) = x^2 + \left(\frac{x^2}{4} - 1\right)^2$. Taking the derivative of f yields $x + \frac{1}{4}x^3$ and setting this equal to 0 leads to the solution, $x = 0$. Therefore, the point closest to $(0, 1)$ is $(0, 0)$. Now let's do it another way. Let's use $y = \frac{x^2}{4}$ to write $x^2 = 4y$. Now for (x, y) on the graph, it follows it is of the form $(\sqrt{4y}, y)$. Therefore, minimize $f(y) = 4y + (y - 1)^2$. Take the derivative to obtain $2 + 2y$ which requires $y = -1$. However, on this graph, y is never negative. What on earth is the problem?

8. Find the dimensions of the rectangle of largest area that can be inscribed in the ellipse, $\frac{x^2}{9} + \frac{y^2}{4} = 1$.
9. A function f , is said to be odd if $f(-x) = -f(x)$ and a function is said to be even if $f(-x) = f(x)$. Show that if f is even, then f' is odd and if f is odd, then f' is even. Sketch the graph of a typical odd function and a typical even function.
10. Find the point on the curve, $y = \sqrt{25 - 2x}$ which is closest to $(0, 0)$.
11. A street is 200 feet long and there are two lights located at the ends of the street. One of the lights is $\frac{1}{8}$ times as bright as the other. Assuming the brightness of light from one of these street lights is proportional to the brightness of the light and the reciprocal of the square of the distance from the light, locate the darkest point on the street.
12. Find the maximum and minimum values for the following functions defined on the given intervals.

(a) $x^3 - 3x^2 + x - 7$, $[0, 4]$	(g) $1 - 2x^2 + x^4$, $[-2, 2]$
(b) $\ln(x^2 - x + 2)$, $[0, 2]$	(h) $\ln(2 - 2x^2 + x^4)$, $[-1, 2]$
(c) $x^3 + 3x$, $[-1, 10]$	(i) $x^2 + 4x - 8$, $[-4, 2]$
(d) $\frac{x^2 + 1 + 3x^3}{3x^2 + 5}$, $[-1, 1]$	(j) $x^2 - 3x + 6$, $[-2, 4]$
(e) $\sin(x^3 - x)$, $[-1, 1]$	(k) $-x^2 + 3x$, $[-4, 2]$
(f) $x^2 - x \tan x$, $[-1, 1]$	(l) $x + \frac{1}{x}$, $(0, \infty)$
13. A cylindrical can is to be constructed to hold 30 cubic inches. The top and bottom of the can are constructed of a material costing one cent per square inch and the sides are constructed of a material costing 2 cents per square inch. Find the minimum cost for such a can.
14. Two positive numbers sum to 8. Find the numbers if their product is to be as large as possible.
15. The ordered pair (x, y) is on the ellipse $x^2 + 4y^2 = 4$. Form the rectangle which has (x, y) as one end of a diagonal and $(0, 0)$ at the other end. Find the rectangle of this sort which has the largest possible area.
16. A rectangle is inscribed in a circle of radius r . Find the formula for the rectangle of this sort which has the largest possible area.

17. A point is picked on the ellipse $x^2 + 4y^2 = 4$ which is in the first quadrant. Then a line tangent to this point is drawn which intersects the x axis at a point x_1 and the y axis at the point y_1 . The area of the triangle formed by the y axis, the x axis, and the line just drawn is thus $\frac{x_1 y_1}{2}$. Out of all possible triangles formed in this way, find the one with smallest area.
18. Find maximum and minimum values if they exist for the function $f(x) = \frac{\ln x}{x}$ for $x > 0$.
19. Describe how you would find the maximum value of the function $f(x) = \frac{\ln x}{2 + \sin x}$ for $x \in (0, 6)$ if it exists. **Hint:** You might want to use a calculator to graph this and get an idea what is going on.
20. A rectangular beam of height h and width w is to be sawed from a circular log of radius 1 foot. Find the dimensions of the strongest such beam assuming the strength is of the form kh^2w . Here k is some constant which depends on the type of wood used.



21. A farmer has 600 feet of fence with which to enclose a rectangular piece of land that borders a river. If he can use the river as one side, what is the largest area that he can enclose.
22. An open box is to be made by cutting out little squares at the corners of a rectangular piece of cardboard which is 20 inches wide and 40 inches long and then folding up the rectangular tabs which result. What is the largest possible volume which can be obtained?
23. A feeding trough is to be made from a rectangular piece of metal which is 3 feet wide and 12 feet long by folding up two rectangular pieces of dimension one foot by 12 feet. What is the best angle for this fold?
24. Find the dimensions of the right circular cone which has the smallest area given the volume is 30π cubic inches. The volume of the right circular cone is $(1/3)\pi r^2 h$ and the area of the cone is $\pi r \sqrt{h^2 + r^2}$.
25. A wire of length 10 inches is cut into two pieces, one of length x and the other of length $10 - x$. One piece is bent into the shape of a square and the other piece is bent into the shape of a circle. Find the two lengths such that the sum of the areas of the circle and the square is as large as possible. What are the lengths if the sum of the two areas is to be as small as possible.
26. A hiker begins to walk to a cabin in a dense forest. He is walking on a road which runs from East to West and the cabin is located exactly one mile north of a point two miles down the road. He walks 5 miles per hour on the road but only 3 miles per hour in the woods. Find the path which will minimize the time it takes for him to get to the cabin.

27. A park ranger needs to get to a fire observation tower which is one mile from a long straight road in a dense forest. The point on the road closest to the observation tower is 10 miles down the road on which the park ranger is standing. Knowing that he can walk at 4 miles per hour on the road but only one mile per hour in the forest, how far down the road should he walk before entering the forest, in order to minimize the travel time?
28. A refinery is on a straight shore line. Oil needs to flow from a mooring place for oil tankers to this refinery. Suppose the mooring place is two miles off shore from a point on the shore 8 miles away from the refinery which is also on the shore and that it costs five times as much to lay pipe under water than above the ground. Describe the most economical route for a pipeline from the mooring place to the refinery.
29. Two hallways, one 5 feet wide and the other 6 feet wide meet. It is desired to carry a ladder horizontally around the corner. What is the longest ladder which can be carried in this way? **Hint:** Consider a line through the inside corner which extends to the opposite walls. The shortest such line will be the length of the longest ladder.
30. A window is to be constructed for the wall of a church which is to consist of a rectangle of height b surmounted by a half circle of radius a . Suppose the total perimeter of the window is to be no more than $4\pi + 8$ feet. Find the dimensions of the window which will admit the most light.
31. * A parabola opens down. The vertex is at the point $(0, a)$ and the parabola intercepts the x axis at the points $(-b, 0)$ and $(b, 0)$. A tangent line to the parabola is drawn in the first quadrant which has the property that the triangle formed by this tangent line and the x and y axes has smallest possible area. Find a relationship between a and b such that the normal line to the point of tangency passes through $(0, 0)$. Also determine what kind of triangle this is.
32. Show that for r a rational number and $y = x^r$, it must be the case that if this function is differentiable, then $y' = rx^{r-1}$. This was shown in more generality, but use the chain rule to verify this directly.
33. Let

$$f(x) = \begin{cases} 1 & \text{if } x \in \mathbb{Q} \\ 0 & \text{if } x \notin \mathbb{Q} \end{cases}$$

Now let $g(x) = x^2 f(x)$. Find where g is continuous and differentiable if anywhere.

34. Use induction to show that for u, v smooth functions,

$$\frac{d^n}{dx^n}(uv) = \sum_{k=0}^n \binom{n}{k} u^{(n-k)} v^{(k)}$$

Here $v^{(k)}$ denotes the k^{th} derivative of v .

5.11 Mean Value Theorem

The mean value theorem is the most important theorem about the derivative of a function of one variable. It pertains only to a real valued function of a real variable. The best versions of many other theorems depend on this fundamental result. The mean value theorem is based on the following special case known as Rolle's theorem¹. It is an existence theorem and like the other existence theorems in analysis, it depends on the completeness axiom. This was only realized in the nineteenth century.

Theorem 5.11.1 Suppose $f : [a, b] \rightarrow \mathbb{R}$ is continuous, $f(a) = f(b)$, and $f : (a, b) \rightarrow \mathbb{R}$ has a derivative at every point of (a, b) . Then there exists $x \in (a, b)$ such that $f'(x) = 0$.

Proof: Suppose first that $f(x) = f(a)$ for all $x \in [a, b]$. Then any $x \in (a, b)$ is a point such that $f'(x) = 0$. If f is not constant, either there exists $y \in (a, b)$ such that $f(y) > f(a)$ or there exists $y \in (a, b)$ such that $f(y) < f(b)$. In the first case, the maximum of f is achieved at some $x \in (a, b)$ and in the second case, the minimum of f is achieved at some $x \in (a, b)$. Either way, Theorem 5.9.2 implies $f'(x) = 0$. ■

The next theorem is known as the Cauchy mean value theorem. It is the best version of this important theorem.

Theorem 5.11.2 Suppose f, g are continuous on $[a, b]$, differentiable on (a, b) . Then there exists $x \in (a, b)$ such that

$$f'(x)(g(b) - g(a)) = g'(x)(f(b) - f(a)).$$

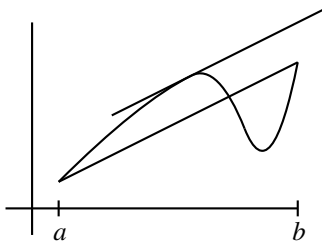
Proof: Let

$$h(x) \equiv f(x)(g(b) - g(a)) - g(x)(f(b) - f(a)).$$

Then letting $x = a$ and then letting $x = b$, a short computation shows $h(a) = h(b)$. Also, h is continuous on $[a, b]$ and differentiable on (a, b) . Therefore Rolle's theorem applies and there exists $x \in (a, b)$ such that

$$h'(x) = f'(x)(g(b) - g(a)) - g'(x)(f(b) - f(a)) = 0. \blacksquare$$

Letting $g(x) = x$, the usual version of the mean value theorem is obtained. Here is the usual picture which describes the theorem.



Corollary 5.11.3 Let f be a continuous real valued function defined on $[a, b]$ and differentiable on (a, b) . Then there exists $x \in (a, b)$ such that $f(b) - f(a) = f'(x)(b - a)$.

¹Rolle is remembered for Rolle's theorem more than his work on diophantine equations. Ironically, he did not like calculus, in particular infinitesimals. These somewhat ill defined ideas were finally expunged from calculus in the nineteenth century when the concept of limits, and completeness of \mathbb{R} were carefully formulated. The notion of infinitesimals can be made precise but this was not the case back then.

Note that $f(a) - f(b) = f'(x)(a - b)$.

Corollary 5.11.4 Suppose $f'(x) = 0$ for all $x \in (a, b)$ where $a \geq -\infty$ and $b \leq \infty$. Then $f(x) = f(y)$ for all $x, y \in (a, b)$. Thus f is a constant.

Proof: If this is not true, there exists x_1 and x_2 such that $f(x_1) \neq f(x_2)$. Then by the mean value theorem, $0 \neq \frac{f(x_1) - f(x_2)}{x_1 - x_2} = f'(z)$ for some z between x_1 and x_2 . This contradicts the hypothesis that $f'(x) = 0$ for all x . ■

Corollary 5.11.5 Suppose $f'(x) > 0$ for all $x \in (a, b)$ where $a \geq -\infty$ and $b \leq \infty$. Then f is strictly increasing on (a, b) . That is, if $x < y$, then $f(x) < f(y)$. If $f'(x) \geq 0$, then f is non-decreasing in the sense that whenever $x < y$ it follows that $f(x) \leq f(y)$. If $f'(x) < 0$ on (a, b) , replace “increasing” with decreasing. If $f'(x) \leq 0$ on (a, b) , replace non-decreasing with non-increasing.

Proof: Let $x < y$. Then by the mean value theorem, there exists $z \in (x, y)$ such that $0 < f'(z) = \frac{f(y) - f(x)}{y - x}$. Since $y > x$, it follows $f(y) > f(x)$ as claimed. Replacing $<$ by \leq in the above equation and repeating the argument gives the second claim. The last claims are shown similarly. ■

Suppose $f(t)$ gives the x coordinate at time t of an object. Then the average velocity on the time interval $[t, t + h]$ equals $\frac{f(t+h) - f(t)}{h}$. By the mean value theorem, this would equal $f'(s)$ for some $s \in (t, t + h)$. Assuming that f' is continuous, when you allow $h \rightarrow 0$, this yields that the instantaneous velocity should be defined as $f'(t)$. The speed is defined as the magnitude of the velocity. Thus the speed is $|f'(t)|$.

5.12 Exercises

1. Sally drives her Saturn over the 110 mile toll road in exactly 1.3 hours. The speed limit on this toll road is 70 miles per hour and the fine for speeding is 10 dollars per mile per hour over the speed limit. How much should Sally pay?
2. Two cars are careening down a freeway in Utah weaving in and out of traffic, which is itself exceeding the speed limit. Car A passes car B and then car B passes car A as the driver makes obscene gestures. This infuriates the driver of car A who passes car B while firing his handgun at the driver of car B. Show there are at least two times when both cars have the same speed. Then show there exists at least one time when they have the same acceleration. The acceleration is the derivative of the velocity.
3. Show the cubic function $f(x) = 5x^3 + 7x - 18$ has only one real zero.
4. Suppose $f(x) = x^7 + |x| + x - 12$. How many solutions are there to the equation, $f(x) = 0$?
5. Let $f(x) = |x - 7| + (x - 7)^2 - 2$ on the interval $[6, 8]$. Then $f(6) = 0 = f(8)$. Does it follow from Rolle's theorem that there exists $c \in (6, 8)$ such that $f'(c) = 0$? Explain your answer.
6. Suppose f and g are differentiable functions defined on \mathbb{R} . Suppose also that it is known that $|f'(x)| > |g'(x)|$ for all x and that $|f'(t)| > 0$ for all t . Show that whenever $x \neq y$, it follows $|f(x) - f(y)| > |g(x) - g(y)|$. **Hint:** Use the Cauchy mean value theorem, Theorem 5.11.2.

7. Show that, like continuous functions, functions which are derivatives have the intermediate value property. This means that if $f'(a) < 0 < f'(b)$ then there exists $x \in (a, b)$ such that $f'(x) = 0$. **Hint:** Argue the minimum value of f occurs at an interior point of $[a, b]$.
8. Find an example of a function which has a derivative at every point but such that the derivative is not everywhere continuous. **Hint:** Consider something involving $x^2 \sin(1/x)$.
9. *Let f be a real continuous function defined on the interval $[0, 1]$. Also suppose $f(0) = 0$ and $f(1) = 1$ and $f'(t)$ exists for all $t \in (0, 1)$. Show there exists n distinct points $\{s_i\}_{i=1}^n$ of the interval such that $\sum_{i=1}^n f'(s_i) = n$. **Hint:** Consider the mean value theorem applied to successive pairs in the following sum. $f(\frac{1}{3}) - f(0) + f(\frac{2}{3}) - f(\frac{1}{3}) + f(1) - f(\frac{2}{3})$
10. *Now suppose $f: [0, 1] \rightarrow \mathbb{R}$ is continuous and differentiable on $(0, 1)$ and $f(0) = 0$ while $f(1) = 1$. Show there are distinct points $\{s_i\}_{i=1}^n \subseteq (0, 1)$ such that

$$\sum_{i=1}^n (f'(s_i))^{-1} = n.$$

Hint: Let $0 = t_0 < t_1 < \dots < t_n = 1$ and pick $x_i \in f^{-1}(t_i)$ such that these x_i are increasing and $x_n = 1, x_0 = 0$. Explain why you can do this. Then argue $t_{i+1} - t_i = f(x_{i+1}) - f(x_i) = f'(s_i)(x_{i+1} - x_i)$ and so $\frac{x_{i+1} - x_i}{t_{i+1} - t_i} = \frac{1}{f'(s_i)}$. Now choose the t_i to be equally spaced.

11. Show that $(x+1)^{3/2} - x^{3/2} > 2$ for all $x \geq 2$. Explain why for n a natural number larger than or equal to 1, there exists a natural number m such that $(n+1)^3 > m^2 > n^3$. **Hint:** Verify directly for $n = 1$ and use the above inequality to take care of the case where $n \geq 2$. This shows that between the cubes of any two natural numbers there is the square of a natural number. This interesting fact was used by Jacobi in 1835 to show a very important theorem in complex analysis.
12. An initial value problem for undamped vibration is

$$\underbrace{y'' + \omega^2 y = 0}_{\text{differential equation}}, \underbrace{y(0) = y_0, y'(0) = y_1}_{\text{initial conditions}}$$

You are looking for a function $y(t)$ which satisfies this equation.

- (a) First show that if you have a complex valued function $z(t)$ satisfying the differential equation, then the real and imaginary parts of z denoted by $\operatorname{Re} z$ and $\operatorname{Im} z$ also solve the differential equation.
- (b) Show that if y_1 and y_2 solve the differential equation, then if C_1, C_2 are arbitrary constants, then $C_1 y_1 + C_2 y_2$ also solves the differential equation.
- (c) Now use Euler's formula in Section 5.6 to show that $z = e^{i\omega t}, z = e^{-i\omega t}$ solve the differential equation. Use the first part to find that $y_1(t) = \sin \omega t$ and $y_2(t) = \cos \omega t$ both solve the above equation.

- (d) Show there exist constants C_1, C_2 such that $C_1 \cos(\omega t) + C_2 \sin(\omega t)$ solve both the differential equation and the initial conditions where y_1 is the real part and y_2 is the imaginary part of z .
- (e) Show that there is only one solution to $y'' + \omega^2 y = 0, y(0) = y_0, y'(0) = y_1$ by assuming there are two. Then the difference of these two would satisfy $y'' + \omega^2 y = 0, y(0) = 0, y'(0) = 0$. Verify the only solution to this is $y = 0$.
13. Find the solution to the differential equation $y'' + 2ay' + b^2y = 0, y(0) = y_0, y'(0) = y_1$ assuming that $b^2 - a^2 > 0$. **Hint:** Re write the equation. Let $z = e^{at}y$. Then show that $z'' + (b^2 - a^2)z = 0$ and $z(0) = y_0, z'(0) = ay_0 + y_1$ use the above problem if $b^2 - a^2 > 0$.

5.13 First and Second Derivative Tests

These tests are sometimes used to determine whether a critical point is a local minimum or a local maximum. First consider the first derivative test.

Theorem 5.13.1 Suppose f is defined near x and is differentiable on

$$(x - \delta, x) \cup (x, x + \delta)$$

and continuous on $[x - \delta, x] \cup [x, x + \delta]$ and suppose $f'(y) < 0$ for $y \in (x - \delta, x)$ and $f'(y) > 0$ for $y \in (x, x + \delta)$. Then x is a local minimum point. That is, for all $y \in (x - \delta, x + \delta)$, $f(x) < f(y)$. If f is defined near x and is differentiable on $(x - \delta, x) \cup (x, x + \delta)$ and continuous on $[x - \delta, x] \cup [x, x + \delta]$ and $f'(y) > 0$ for $y \in (x - \delta, x)$ and $f'(y) < 0$ for $y \in (x, x + \delta)$. Then x is a local maximum point. That is, for all $y \in (x - \delta, x + \delta)$, $f(x) > f(y)$. You can replace all strict inequalities with the corresponding less than or equal or greater than or equal.

Proof: This follows right away from the mean value theorem. Suppose the first case. The proof of the second is exactly the same. Say $y \in (x - \delta, x)$. Then by the mean value theorem, $f(y) - f(x) = f'(t)(y - x) > 0$ because both $f'(t), (y - x) < 0$. If $y \in (x, x + \delta)$ then $f(y) - f(x) = f'(t)(y - x) > 0$ because by assumption both terms $f'(t)$ and $(y - x)$ are positive. In the second case it works the same way with the inequalities all turned around. It all works the same way if you replace $<$ with \leq and $>$ with \geq . ■

Now it is time for the second derivative test. It is an inferior sort of thing since it does not always work.

Lemma 5.13.2 Suppose f has a continuous derivative near a point x and also $f'(x) > 0$. Then there exists $\delta > 0$ such that f is increasing on $(x - \delta, x + \delta)$. If $f'(x) < 0$, then there exists $\delta > 0$ such that f is decreasing on $(x - \delta, x + \delta)$.

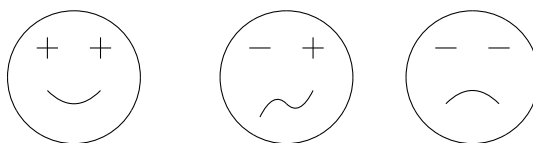
Proof: If $f'(x) > 0$ then there exists $\delta > 0$ such that $f'(y) > 0$ for $y \in (x - \delta, x + \delta)$ thanks to continuity of f' . Then by the mean value theorem, f is increasing on this interval. Indeed if x, \hat{x} are in $(x - \delta, x + \delta)$ with $x < \hat{x}$, then $f(\hat{x}) - f(x) = f'(y)(\hat{x} - x) > 0$. It works the same way if $f'(x) < 0$ except the inequalities are turned around. ■

Theorem 5.13.3 Suppose f , defined on an open interval, has continuous first and second derivatives near x and $f'(x) = 0$. Then

1. If $f''(x) > 0$, then x is a local minimum.
2. If $f''(x) < 0$ then x is a local maximum.
3. If $f''(x) = 0$ the test fails.

Proof: Consider case 1. By Lemma 5.13.2, there is an interval $(x - \delta, x + \delta)$ on which f' is strictly increasing. Thus $f'(y) > 0$ for $y > x$ and $f'(y) < 0$ for $y < x$. By Theorem 5.13.1, x is a local minimum. Part 2. is similar. Just turn around the inequalities. To see the test fails if $f''(x) = 0$, consider $f(x) = x^4$, $f(x) = -x^4$ and $f(x) = x^3$. Each has $f'(0) = 0$ the first is a local minimum at 0, the second has a local maximum at 0 and the third has neither a local minimum nor a local maximum at 0. ■

The following picture may help remember. The idea is that if the second derivative is positive, the shape of the curve is a smile. If the second derivative is negative, the graph is a frown. If it is neither positive nor negative, you don't know what the graph is. It could be smiling, frowning, or neither.



Example 5.13.4 Consider $f(x) \equiv x^4 - x^3$. Find and classify the critical points.

To find critical points, take the derivative and set equal to 0. Thus $4x^3 - 3x^2 = 0$ and so $x = 0, 0, \frac{3}{4}$. The second derivative is $12x^2 - 6x$. This equals 0 when $x = 0$ so the second derivative fails. However, you could look at the first derivative. It is negative if x is small and positive and it is also negative if x is small and negative. Therefore, the function is decreasing near 0. As to the other critical point, $12\left(\frac{3}{4}\right)^2 - 6\left(\frac{3}{4}\right) = \frac{9}{4}$ which is positive and so this critical point is a local minimum.

5.14 Exercises

1. For $1 \geq x \geq 0$, and $p \geq 1$, show that $(1 - x)^p \geq 1 - px$. **Hint:** This can be done using the mean value theorem. Define $f(x) \equiv (1 - x)^p - 1 + px$ and show that $f(0) = 0$ while $f'(x) \geq 0$ for all $x \in (0, 1)$.
2. The graph of a function $y = f(x)$ is said to be “convex” if whenever $t \in [0, 1]$,

$$f(x + t(y - x)) \leq (1 - t)f(x) + tf(y)$$

Show that if f is twice differentiable on an open interval, (a, b) and $f''(x) > 0$ for all x , then the graph of f is convex.

3. Suppose you have a function f which has two derivatives. Suppose $f'' > 0$. Give a sketch of the graph of f . In particular, show that the overall shape of the function is that it curves up. See the next problem.

4. Show that if the graph of a function f defined on an interval (a, b) is convex, then if f' exists on (a, b) , it must be the case that f' is a non decreasing function. Note you do not know the second derivative exists.
5. Convex functions defined in Problem 2 have a very interesting property. Suppose $\{a_i\}_{i=1}^n$ are all nonnegative, sum to 1, and suppose ϕ is a convex function defined on \mathbb{R} . Then

$$\phi\left(\sum_{k=1}^n a_k x_k\right) \leq \sum_{k=1}^n a_k \phi(x_k).$$

Verify this interesting inequality.

6. Find all critical points of the function $f(x) = \frac{1}{4}x^4 - 2x^3 + \frac{11}{2}x^2 - 6x$. Classify each critical point according to whether it is a local minimum, maximum, or neither.
7. Let $f(x) = \cos(x) + \sin(x)$. Describe the critical points and classify these. **Hint:** You might try writing as $\sqrt{2}\cos(x - \frac{\pi}{4})$.
8. Let $f(x) \equiv x^4 - 6x^3 + 12x^2 - 10x + 3$. Find and classify the critical points.

5.15 Taylor Series Approximations

One of the really nice applications of the derivative is to the approximation of functions like $\sin(x)$ with a polynomial. The reason this is so nice is that it is easy to compute the value of a polynomial at various points. If someone asks for $\sin(.1)$, how do you find it? You may say you look on your calculator, but how does it find it? In this section, approximation with polynomials will be discussed. The main result is the following theorem. A version is due to Lagrange, about 1790.²

Theorem 5.15.1 Suppose f has $n+1$ derivatives (That is $f^{(n+1)}(t)$ exists for $t \in (a, b)$) on an interval (a, b) and let $c \in (a, b)$. Then if $x \in (a, b)$, there exists ξ between c and x such that

$$f(x) = f(c) + \sum_{k=1}^n \frac{f^{(k)}(c)}{k!} (x-c)^k + \frac{f^{(n+1)}(\xi)}{(n+1)!} (x-c)^{n+1}.$$

(In this formula, the symbol $\sum_{k=1}^0 a_k$ will denote the number 0.)

²Joseph-Louis Lagrange, 25 January 1736 - 10 April 1813 was very important in the development of mathematics. He was actually Italian but lived in France. The above is his French name. With Euler, he invented the calculus of variations. The Euler Lagrange equations are due to them. He wrote *Mécanique analytique* an important work on mechanics. Lagrange was able to describe analytically the motion of a spinning top using so called Lagrangian mechanics which is an amazing method used to obtain differential equations of motion. He also made major contributions to number theory and astronomy. He was among the group of scientists who provided us with the metric system. Like Laplace, another important figure in the development of mathematics, and in contrast to many earlier mathematicians, Lagrange was not particularly concerned with theology, a typical attitude for this time.

This period of time, sometimes called the enlightenment, saw major intellectual achievements in virtually every human concern. By contrast, in frontier America, believers in magic attempted to mollify guardian spirits through suitable rituals to prevent buried treasure from slipping further into the earth before they could get it. Some also created bizarre religious cults.

Proof: There exists K such that

$$f(x) - \left(f(c) + \sum_{k=1}^n \frac{f^{(k)}(c)}{k!} (x-c)^k + K(x-c)^{n+1} \right) = 0 \quad (5.6)$$

In fact, solving for K ,

$$K = \frac{-f(x) + \left(f(c) + \sum_{k=1}^n \frac{f^{(k)}(c)}{k!} (x-c)^k \right)}{(x-c)^{n+1}}.$$

Now define $F(t)$ for t in the closed interval determined by x and c by

$$F(t) \equiv f(x) - \left(f(t) + \sum_{k=1}^n \frac{f^{(k)}(t)}{k!} (x-t)^k + K(x-t)^{n+1} \right).$$

The c in 5.6 got replaced by t .

Therefore, $F(c) = 0$ and also $F(x) = 0$. Then

$$\begin{aligned} F'(t) &= - \left(f'(t) - \left(\sum_{k=1}^n \frac{f^{(k)}(t)}{k!} k (x-t)^{k-1} - \sum_{k=1}^n \frac{f^{(k+1)}(t)}{k!} (x-t)^k + K(n+1)(x-t)^n \right) \right) \\ &= - \left(f'(t) - \left(\sum_{k=0}^{n-1} \frac{f^{(k+1)}(t)}{k!} (x-t)^k - \sum_{k=1}^n \frac{f^{(k+1)}(t)}{k!} (x-t)^k + K(n+1)(x-t)^n \right) \right) \\ &= - \left(f'(t) - \left(f'(t) - f^{(n+1)}(t)(x-t)^n + K(n+1)(x-t)^n \right) \right) \\ &= -f'(t) + f'(t) - f^{(n+1)}(t)(x-t)^n + K(n+1)(x-t)^n \\ &= -f^{(n+1)}(t) \frac{1}{n!} (x-t)^n + K(n+1)(x-t)^n \end{aligned}$$

By the mean value theorem or Rolle's theorem, there exists ξ between x and c such that $F'(\xi) = 0$. Therefore,

$$-f^{(n+1)}(\xi) \frac{1}{n!} (x-\xi)^n + K(n+1)(x-\xi)^n = 0$$

and so $K(n+1) = f^{(n+1)}(\xi) \frac{1}{n!}$, $K = \frac{f^{(n+1)}(\xi)}{(n+1)!}$ ■

The term $\frac{f^{(n+1)}(\xi)}{(n+1)!} (x-c)^{n+1}$, is called the remainder and this particular form of the remainder is called the Lagrange form of the remainder.

Note how the approximations depend on the derivatives evaluated at c and the part of the approximation before the remainder is called the Taylor series approximation for f expanded about c .

Example 5.15.2 Find the Taylor series for e^x expanded about 0.

In this case, all derivatives are e^x and so $f^{(n)}(0)$, the n^{th} derivative evaluated at 0, is always 1. Therefore,

$$e^x = \sum_{k=0}^n \frac{x^k}{k!} + \frac{e^\xi x^{n+1}}{(n+1)!}, \text{ some } \xi \text{ between 0 and } x. \quad (5.7)$$

Example 5.15.3 Show $e^x = \lim_{n \rightarrow \infty} \sum_{k=0}^n \frac{x^k}{k!}$.

This is pretty easy if $|x| \leq 1$. This is because $\left| e^x - \sum_{k=0}^n \frac{x^k}{k!} \right| \leq \frac{e^\xi}{(n+1)!} \leq \frac{e}{(n+1)!}$. Clearly the term at the end converges to 0 as $n \rightarrow \infty$. Now suppose you have arbitrary x . This works out exactly the same for arbitrary x if it can be shown that $\lim_{n \rightarrow \infty} \frac{e^{|x|} |x|^{n+1}}{(n+1)!} = 0$. This is because $x \rightarrow e^x$ is an increasing function. It suffices to show that for $r > 0$, $\lim_{n \rightarrow \infty} \frac{r^n}{n!} = 0$. Why would this be so? It is because $\frac{(r^{n+1}/(n+1)!)}{r^n/n!} = \frac{r}{n+1} < \frac{1}{2}$ for all n large enough. Say this happens for all $n \geq N$. Then, letting $a_n = \frac{r^n}{n!}$ to save space, this has shown that there is some N such that $a_{k+1}/a_k \leq \frac{1}{2}$ for all $k \geq N$. Thus, for $k > N$,

$$a_{k+1} \leq \frac{1}{2} a_k \leq \frac{1}{2^2} a_{k-1} \leq \cdots \leq \frac{1}{2^{k-N}} a_N = \frac{1}{2^k} (a_N 2^N)$$

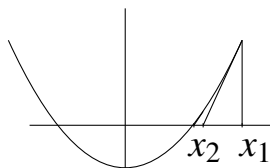
clearly $\lim_{k \rightarrow \infty} \frac{1}{2^k} = 0$ and so this shows the desired result that $\lim_{n \rightarrow \infty} \frac{r^n}{n!} = 0$.

Exercise 5.15.4 What is e to several decimal places?

From the above, $\left| e - \sum_{k=0}^n \frac{1}{k!} \right| \leq \frac{e}{(n+1)!}$. One can easily see that $\ln 3 > 1$ directly from the definition of \ln . Therefore, $3 > e$. Also $\frac{3}{10!} \leq 10^{-6}$. It follows that $\sum_{k=0}^{10} \frac{1}{k!}$ is within 10^{-6} of e . Therefore, $\sum_{k=0}^{10} \frac{1}{k!} = 2.7182818$ is within 10^{-6} of e .

5.16 Exercises

1. Let f have n derivatives on an open interval containing c . Suppose you desire to approximate $f(x)$ with a polynomial $p_n(x) = \sum_{k=0}^n a_k (x-c)^k$ such that both p_n and f have the same first n derivatives at c . Show that it must be the case that $a_k = \frac{f^{(k)}(c)}{k!}$.
2. Show that $\sin(x) = \sum_{k=0}^n (-1)^k \frac{x^{2k+1}}{(2k+1)!} + \frac{\sin^{(2n+2)}(\xi)x^{2n+2}}{(2n+2)!}$ for some ξ between 0 and x . Find $\sin(.1)$ to a few decimal places and estimate how close your approximation is using the remainder term.
3. Show that $\cos(x) = \sum_{k=0}^n (-1)^k \frac{x^{2k}}{(2k)!} + \frac{\cos^{(2n+2)}(\xi)x^{2n+1}}{(2n+1)!}$ for some ξ between 0 and x . Find $\cos(.1)$ to a few decimal places and estimate how close your approximation is using the remainder term.
4. Explain why, for $|x| \leq 1$, $\cos(x) = \lim_{n \rightarrow \infty} \sum_{k=0}^n (-1)^k \frac{x^{2k}}{(2k)!}$.
5. Suppose you want to find a function y such that $y'(x) + xy(x) = \sin(x)$ and $y(0) = 1$. This is called an initial value problem for y . Find a polynomial of degree 3 which will approximate the solution to this equation in the sense that the first three derivatives of both y and the polynomial coincide at $x = 0$, assuming there is such a solution y . **Hint:** Use Problem 1 and the differential equation to determine this polynomial.
6. The following is the graph of a function and there are two points indicated $(x_1, 0)$ and $(x_2, 0)$, the latter coming from the intersection of the tangent line to the graph of the function at $(x_1, f(x_1))$ and the x axis as shown.



Determine a formula for x_2 in terms of the function and its derivative evaluated at x_1 . The idea is that x_2 is a better approximation to a solution to $f(x) = 0$ than x_1 . Now describe an iterative procedure which hopefully will yield a sequence of approximate solutions to $f(x) = 0$ which converges to a solution to this equation. If you do it right, it is called the Newton Raphson procedure.

7. Use the above Newton Raphson procedure to find $\sqrt{3}$ valid to four decimal places.
8. Consider the function $y = x^{1/3}$ which has a zero at $x = 0$. Show that the above Newton Raphson method will not work for this example. Is there some condition which will cause the above procedure to work?
9. If x is small and positive, explain why $\tan x - x > 0$. **Hint:** This amounts to showing that $\sin x > x \cos(x)$. Now use Taylor series approximations.

5.17 L'Hôpital's Rule

There is an interesting rule which is often useful for evaluating difficult limits. This is called L'Hôpital's³ rule. The best versions of this rule are based on the Cauchy Mean value theorem, Theorem 5.11.2 on Page 147.

Theorem 5.17.1 Let $[a, b] \subseteq [-\infty, \infty]$ and suppose f, g are functions which satisfy,

$$\lim_{x \rightarrow b-} f(x) = \lim_{x \rightarrow b-} g(x) = 0, \quad (5.8)$$

and f' and g' exist on (a, b) with $g'(x) \neq 0$ on (a, b) . Suppose also that

$$\lim_{x \rightarrow b-} \frac{f'(x)}{g'(x)} = L. \quad (5.9)$$

Then

$$\lim_{x \rightarrow b-} \frac{f(x)}{g(x)} = L. \quad (5.10)$$

Proof: By the definition of limit and 5.9 there exists $c < b$ such that if $t > c$, then

$$\left| \frac{f'(t)}{g'(t)} - L \right| < \frac{\varepsilon}{2}.$$

³L'Hôpital published the first calculus book in 1696. This rule, named after him, appeared in this book. The rule was actually due to Bernoulli who had been L'Hôpital's teacher. L'Hôpital did not claim the rule as his own but Bernoulli accused him of plagiarism. Nevertheless, this rule has become known as L'Hôpital's rule ever since. There was entirely too much squabbling about who originated various ideas during this period of time. The version of the rule presented here is superior to what was discovered by Bernoulli and depends on the Cauchy mean value theorem which was found over 100 years after the time of L'Hôpital. Cauchy often saw things which were both significant and unobserved by all the others before him. In addition to this, he invented whole new parts of mathematics such as complex analysis and made significant contributions to mechanics and algebra.

Now pick x, y such that $c < x < y < b$. By the Cauchy mean value theorem, there exists $t \in (x, y)$ such that

$$g'(t)(f(x) - f(y)) = f'(t)(g(x) - g(y)).$$

Since $g'(s) \neq 0$ for all $s \in (a, b)$ it follows from the mean value theorem $g(x) - g(y) \neq 0$. Therefore,

$$\frac{f'(t)}{g'(t)} = \frac{f(x) - f(y)}{g(x) - g(y)}$$

and so, since $t > c$,

$$\left| \frac{f'(t)}{g'(t)} - L \right| = \left| \frac{f(x) - f(y)}{g(x) - g(y)} - L \right| < \frac{\varepsilon}{2}.$$

Now taking $\lim_{y \rightarrow b-}$,

$$\left| \frac{f(x)}{g(x)} - L \right| \leq \frac{\varepsilon}{2} < \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, this shows 5.10. ■

The following corollary is proved in the same way.

Corollary 5.17.2 *Let $[a, b] \subseteq [-\infty, \infty]$ and suppose f, g are functions which satisfy,*

$$\lim_{x \rightarrow a+} f(x) = \lim_{x \rightarrow a+} g(x) = 0, \quad (5.11)$$

and f' and g' exist on (a, b) with $g'(x) \neq 0$ on (a, b) . Suppose also that

$$\lim_{x \rightarrow a+} \frac{f'(x)}{g'(x)} = L. \quad (5.12)$$

Then

$$\lim_{x \rightarrow a+} \frac{f(x)}{g(x)} = L. \quad (5.13)$$

Here is a simple example which illustrates the use of this rule.

Example 5.17.3 Find $\lim_{x \rightarrow 0} \frac{5x + \sin 3x}{\tan 7x}$.

The conditions of L'Hôpital's rule are satisfied because the numerator and denominator both converge to 0 and the derivative of the denominator is nonzero for x close to 0. Therefore, if the limit of the quotient of the derivatives exists, it will equal the limit of the original function. Thus,

$$\lim_{x \rightarrow 0} \frac{5x + \sin 3x}{\tan 7x} = \lim_{x \rightarrow 0} \frac{5 + 3 \cos 3x}{7 \sec^2(7x)} = \frac{8}{7}.$$

Sometimes you have to use L'Hôpital's rule more than once.

Example 5.17.4 Find $\lim_{x \rightarrow 0} \frac{\sin x - x}{x^3}$.

Note that $\lim_{x \rightarrow 0} (\sin x - x) = 0$ and $\lim_{x \rightarrow 0} x^3 = 0$. Also, the derivative of the denominator is nonzero for x close to 0. Therefore, if $\lim_{x \rightarrow 0} \frac{\cos x - 1}{3x^2}$ exists and equals L , it will follow from L'Hôpital's rule that the original limit exists and equals L . However, $\lim_{x \rightarrow 0} (\cos x - 1) = 0$ and $\lim_{x \rightarrow 0} 3x^2 = 0$ so L'Hôpital's rule can be applied again to consider $\lim_{x \rightarrow 0} \frac{-\sin x}{6x}$. From L'Hôpital's rule, if this limit exists and equals L , it will follow that $\lim_{x \rightarrow 0} \frac{\cos x - 1}{3x^2} = L$ and consequently $\lim_{x \rightarrow 0} \frac{\sin x - x}{x^3} = L$. But, $\lim_{x \rightarrow 0} \frac{-\sin x}{6x} = \lim_{x \rightarrow 0} \left(\frac{-1}{6} \right) \frac{\sin x}{x} = \frac{-1}{6}$. Therefore, by L'Hôpital's rule, $\lim_{x \rightarrow 0} \frac{\sin x - x}{x^3} = \frac{-1}{6}$.

Warning 5.17.5 *Always check assumptions of L'Hôpital's rule before using it.*

Example 5.17.6 Find $\lim_{x \rightarrow 0^+} \frac{\cos 2x}{x}$.

The numerator becomes close to 1 and the denominator gets close to 0. Therefore, the assumptions of L'Hôpital's rule do not hold and so it does not apply. In fact there is no limit unless you define the limit to equal $+\infty$. Now let's try to use the conclusion of L'Hôpital's rule even though the conditions for using this rule are not verified. Take the derivative of the numerator and the denominator which yields $\frac{-2\sin 2x}{1}$, an expression whose limit as $x \rightarrow 0^+$ equals 0. This is a good illustration of the above warning.

Some people get the unfortunate idea that one can find limits by doing experiments with a calculator. If the limit is taken as x gets close to 0, these people think one can find the limit by evaluating the function at values of x which are closer and closer to 0. Theoretically, this should work although you have no way of knowing how small you need to take x to get a good estimate of the limit. In practice, the procedure may fail miserably.

Example 5.17.7 Find $\lim_{x \rightarrow 0} \frac{\ln|1+x^{10}|}{x^{10}}$.

This limit equals $\lim_{y \rightarrow 0} \frac{\ln|1+y|}{y} = \lim_{y \rightarrow 0} \frac{\left(\frac{1}{1+y}\right)}{1} = 1$ where L'Hôpital's rule has been used. This is an amusing example. You should plug .001 in to the function $\frac{\ln|1+x^{10}|}{x^{10}}$ and see what your calculator or computer gives you. If it is like mine, it will give 0 and will keep on returning the answer of 0 for smaller numbers than .001. This illustrates the folly of trying to compute limits through calculator or computer experiments. Indeed, you could say that a calculator is as useful for understanding limits as a bicycle is for swimming. Those who say otherwise are either guilty of ignorance or dishonesty.

There is another form of L'Hôpital's rule in which

$$\lim_{x \rightarrow b^-} f(x) = \pm\infty \text{ and } \lim_{x \rightarrow b^-} g(x) = \pm\infty.$$

Theorem 5.17.8 Let $[a, b] \subseteq [-\infty, \infty]$ and suppose f, g are functions which satisfy,

$$\lim_{x \rightarrow b^-} f(x) = \pm\infty \text{ and } \lim_{x \rightarrow b^-} g(x) = \pm\infty, \quad (5.14)$$

and f' and g' exist on (a, b) with $g'(x) \neq 0$ on (a, b) . Suppose also

$$\lim_{x \rightarrow b^-} \frac{f'(x)}{g'(x)} = L. \quad (5.15)$$

Then

$$\lim_{x \rightarrow b^-} \frac{f(x)}{g(x)} = L. \quad (5.16)$$

Proof: By the definition of limit and 5.15 there exists $c < b$ such that if $t > c$, then

$$\left| \frac{f'(t)}{g'(t)} - L \right| < \frac{\varepsilon}{2}.$$

Now pick x, y such that $c < x < y < b$. By the Cauchy mean value theorem, there exists $t \in (x, y)$ such that

$$g'(t)(f(x) - f(y)) = f'(t)(g(x) - g(y)).$$

Since $g'(s) \neq 0$ on (a, b) , it follows from mean value theorem $g(x) - g(y) \neq 0$. Therefore,

$$\frac{f'(t)}{g'(t)} = \frac{f(x) - f(y)}{g(x) - g(y)}$$

and so, since $t > c$,

$$\left| \frac{f(x) - f(y)}{g(x) - g(y)} - L \right| < \frac{\varepsilon}{2}.$$

Now this implies

$$\left| \frac{f(y)}{g(y)} \frac{\left(\frac{f(x)}{f(y)} - 1\right)}{\left(\frac{g(x)}{g(y)} - 1\right)} - L \right| < \frac{\varepsilon}{2}$$

where for all y large enough, both $\frac{f(x)}{f(y)} - 1$ and $\frac{g(x)}{g(y)} - 1$ are not equal to zero. Then

$$\left| \frac{f(y)}{g(y)} - L \frac{\left(\frac{g(x)}{g(y)} - 1\right)}{\left(\frac{f(x)}{f(y)} - 1\right)} \right| < \frac{\varepsilon}{2} \left| \frac{\left(\frac{g(x)}{g(y)} - 1\right)}{\left(\frac{f(x)}{f(y)} - 1\right)} \right|.$$

Therefore, for y large enough,

$$\left| \frac{f(y)}{g(y)} - L \right| \leq \left| L - L \frac{\left(\frac{g(x)}{g(y)} - 1\right)}{\left(\frac{f(x)}{f(y)} - 1\right)} \right| + \frac{\varepsilon}{2} \left| \frac{\left(\frac{g(x)}{g(y)} - 1\right)}{\left(\frac{f(x)}{f(y)} - 1\right)} \right| < \varepsilon$$

due to the assumption 5.14 which implies $\lim_{y \rightarrow b-} \frac{\left(\frac{g(x)}{g(y)} - 1\right)}{\left(\frac{f(x)}{f(y)} - 1\right)} = 1$. Therefore, whenever y is large enough, $\left| \frac{f(y)}{g(y)} - L \right| < \varepsilon$ and this is what is meant by 5.16. ■

As before, there is no essential difference between the proof in the case where $x \rightarrow b-$ and the proof when $x \rightarrow a+$. This observation is stated as the next corollary.

Corollary 5.17.9 *Let $[a, b] \subseteq [-\infty, \infty]$ and suppose f, g are functions which satisfy,*

$$\lim_{x \rightarrow a+} f(x) = \pm\infty \text{ and } \lim_{x \rightarrow a+} g(x) = \pm\infty, \quad (5.17)$$

and f' and g' exist on (a, b) with $g'(x) \neq 0$ on (a, b) . Suppose also that

$$\lim_{x \rightarrow a+} \frac{f'(x)}{g'(x)} = L. \quad (5.18)$$

Then

$$\lim_{x \rightarrow a+} \frac{f(x)}{g(x)} = L. \quad (5.19)$$

Theorems 5.17.1 5.17.8 and Corollaries 5.17.2 and 5.17.9 will each be referred to as L'Hôpital's rule from now on. Theorem 5.17.1 and Corollary 5.17.2 involve the notion of indeterminate forms of the form $\frac{0}{0}$. Please do not think any meaning is being assigned to the nonsense expression $\frac{0}{0}$. It is just a symbol to help remember the sort of thing described by Theorem 5.17.1 and Corollary 5.17.2. Theorem 5.17.8 and Corollary 5.17.9 deal with indeterminate forms which are of the form $\frac{\pm\infty}{\infty}$. Again, this is just a symbol which is helpful in remembering the sort of thing being considered. There are other indeterminate forms which can be reduced to these forms just discussed. Don't ever try to assign meaning to such symbols.

Example 5.17.10 Find $\lim_{y \rightarrow \infty} \left(1 + \frac{x}{y}\right)^y$.

It is good to first see why this is called an indeterminate form. One might think that as $y \rightarrow \infty$, it follows $x/y \rightarrow 0$ and so $1 + \frac{x}{y} \rightarrow 1$. Now 1 raised to anything is 1 and so it would seem this limit should equal 1. On the other hand, if $x > 0$, $1 + \frac{x}{y} > 1$ and a number raised to higher and higher powers should approach ∞ . It really isn't clear what this limit should be. It is an indeterminate form which can be described as 1^∞ . By definition,

$$\left(1 + \frac{x}{y}\right)^y = \exp\left(y \ln\left(1 + \frac{x}{y}\right)\right).$$

Now using L'Hôpital's rule,

$$\begin{aligned} \lim_{y \rightarrow \infty} y \ln\left(1 + \frac{x}{y}\right) &= \lim_{y \rightarrow \infty} \frac{\ln\left(1 + \frac{x}{y}\right)}{1/y} = \lim_{y \rightarrow \infty} \frac{\frac{1}{1+(x/y)} (-x/y^2)}{(-1/y^2)} \\ &= \lim_{y \rightarrow \infty} \frac{x}{1 + (x/y)} = x \end{aligned}$$

Therefore, $\lim_{y \rightarrow \infty} y \ln\left(1 + \frac{x}{y}\right) = x$. Since \exp is continuous, it follows

$$\lim_{y \rightarrow \infty} \left(1 + \frac{x}{y}\right)^y = \lim_{y \rightarrow \infty} \exp\left(y \ln\left(1 + \frac{x}{y}\right)\right) = e^x.$$

5.18 Interest Compounded Continuously

Suppose you put money in the bank and it accrues interest at the rate of r per payment period. These terms need a little explanation. If the payment period is one month, and you started with \$100 then the amount at the end of one month would equal $100(1+r) = 100 + 100r$. In this the second term is the interest and the first is called the principal. Now you have $100(1+r)$ in the bank. This becomes the new principal. How much will you have at the end of the second month? By analogy to what was just done it would equal

$$100(1+r) + 100(1+r)r = 100(1+r)^2.$$

In general, the amount you would have at the end of n months is $100(1+r)^n$.

When a bank says they offer 6% compounded monthly, this means r , the rate per payment period equals .06/12. Consider the problem of a rate of r per year and compounding the interest n times a year and letting n increase without bound. This is what is meant by

compounding continuously. The interest rate per payment period is then r/n and the number of payment periods after time t years is approximately tn . From the above the amount in the account after t years is

$$P \left(1 + \frac{r}{n} \right)^{nt} \quad (5.20)$$

Recall from Example 5.17.10 that $\lim_{y \rightarrow \infty} \left(1 + \frac{x}{y} \right)^y = e^x$. The expression in 5.20 can be written as $P \left[\left(1 + \frac{r}{n} \right)^{nt} \right]$ and so, taking the limit as $n \rightarrow \infty$, you get $Pe^{rt} = A$. This shows how to compound interest continuously.

Example 5.18.1 Suppose you have \$100 and you put it in a savings account which pays 6% compounded continuously. How much will you have at the end of 4 years?

From the above discussion, this would be $100e^{(.06)4} = 127.12$. Thus, in 4 years, you would gain interest of about \$27.

5.19 Exercises

1. Find the limits.

- (a) $\lim_{x \rightarrow 0} \frac{3x-4 \sin 3x}{\tan 3x}$
- (b) $\lim_{x \rightarrow \frac{\pi}{2}} (\tan x)^{x-(\pi/2)}$
- (c) $\lim_{x \rightarrow 1} \frac{\arctan(4x-4)}{\arcsin(4x-4)}$
- (d) $\lim_{x \rightarrow 0} \frac{\arctan 3x-3x}{x^3}$
- (e) $\lim_{x \rightarrow 0+} \frac{9^{\sec x}-1}{3^{\sec x}-1}$
- (f) $\lim_{x \rightarrow 0} \frac{3x+\sin 4x}{\tan 2x}$
- (g) $\lim_{x \rightarrow \pi/2} \frac{\ln(\sin x)}{x-(\pi/2)}$
- (h) $\lim_{x \rightarrow 0} \frac{\cosh 2x-1}{x^2}$
- (i) $\lim_{x \rightarrow 0} \frac{-\arctan x+x}{x^3}$
- (j) $\lim_{x \rightarrow 0} \frac{x^8 \sin \frac{1}{x}}{\sin 3x}$

- (k) $\lim_{x \rightarrow \infty} (1+5^x)^{\frac{2}{x}}$
- (l) $\lim_{x \rightarrow 0} \frac{-2x+3 \sin x}{x}$
- (m) $\lim_{x \rightarrow 1} \frac{\ln(\cos(x-1))}{(x-1)^2}$
- (n) $\lim_{x \rightarrow 0+} \sin^{\frac{1}{x}} x$
- (o) $\lim_{x \rightarrow 0} (\csc 5x - \cot 5x)$
- (p) $\lim_{x \rightarrow 0+} \frac{3^{\sin x}-1}{2^{\sin x}-1}$
- (q) $\lim_{x \rightarrow 0+} (4x)^{x^2}$
- (r) $\lim_{x \rightarrow \infty} \frac{x^{10}}{(1.01)^x}$
- (s) $\lim_{x \rightarrow 0} (\cos 4x)^{(1/x^2)}$

2. Find the following limits.

- (a) $\lim_{x \rightarrow 0+} \frac{1-\sqrt{\cos 2x}}{\sin^4(4\sqrt{x})}$
- (b) $\lim_{x \rightarrow 0} \frac{2^{x^2}-2^{5x}}{\sin\left(\frac{x^2}{5}\right)-\sin(3x)}$
- (c) $\lim_{n \rightarrow \infty} n \left(\sqrt[n]{7} - 1 \right)$
- (d) $\lim_{x \rightarrow \infty} \left(\frac{3x+2}{5x-9} \right)^{x^2}$
- (e) $\lim_{x \rightarrow \infty} \left(\frac{3x+2}{5x-9} \right)^{1/x}$
- (f) $\lim_{n \rightarrow \infty} \left(\cos \frac{2x}{\sqrt{n}} \right)^n$
- (g) $\lim_{n \rightarrow \infty} \left(\cos \frac{2x}{\sqrt{5n}} \right)^n$
- (h) $\lim_{x \rightarrow 3} \frac{x^x-27}{x-3}$
- (i) $\lim_{n \rightarrow \infty} \cos \left(\pi \frac{\sqrt{4n^2+13n}}{n} \right)$

$$\begin{aligned}
 \text{(j)} \quad \lim_{x \rightarrow \infty} \left(\frac{\sqrt[3]{x^3 + 7x^2}}{-\sqrt{x^2 - 11x}} \right). & \quad \text{(m)} \quad \lim_{x \rightarrow \infty} \left(\frac{5x^2 + 7}{2x^2 - 11} \right)^{\frac{\sqrt{\ln x}}{1-x}}. \\
 \text{(k)} \quad \lim_{x \rightarrow \infty} \left(\frac{\sqrt[5]{x^5 + 7x^4}}{-\sqrt[3]{x^3 - 11x^2}} \right). & \quad \text{(n)} \quad \lim_{x \rightarrow 0+} \frac{\ln(e^{2x^2} + 7\sqrt{x})}{\sinh(\sqrt{x})}. \\
 \text{(l)} \quad \lim_{x \rightarrow \infty} \left(\frac{5x^2 + 7}{2x^2 - 11} \right)^{\frac{x}{1-x}}. & \quad \text{(o)} \quad \lim_{x \rightarrow 0+} \frac{\sqrt[7]{x} - \sqrt[5]{x}}{\sqrt[9]{x} - \sqrt[11]{x}}.
 \end{aligned}$$

3. Find the following limits.

$$\begin{aligned}
 \text{(a)} \quad \lim_{x \rightarrow 0+} (1 + 3x)^{\cot 2x} & \quad \text{(h)} \quad \lim_{x \rightarrow 0} \left(\frac{1}{x} - \cot(x) \right) \\
 \text{(b)} \quad \lim_{x \rightarrow 0} \frac{\sin x - x}{x^2} = 0 & \quad \text{(i)} \quad \lim_{x \rightarrow 0} \frac{\cos(\sin x) - 1}{x^2} \\
 \text{(c)} \quad \lim_{x \rightarrow 0} \frac{\sin x - x}{x^3} & \quad \text{(j)} \quad \lim_{x \rightarrow \infty} \left(x^2 (4x^4 + 7)^{1/2} - 2x^4 \right) \\
 \text{(d)} \quad \lim_{x \rightarrow 0} \frac{\tan(\sin x) - \sin(\tan x)}{x^7} & \quad \text{(k)} \quad \lim_{x \rightarrow 0} \frac{\cos(x) - \cos(4x)}{\tan(x^2)} \\
 \text{(e)} \quad \lim_{x \rightarrow 0} \frac{\tan(\sin 2x) - \sin(\tan 2x)}{x^7} & \quad \text{(l)} \quad \lim_{x \rightarrow 0} \frac{\arctan(3x)}{x} \\
 \text{(f)} \quad \lim_{x \rightarrow 0} \frac{\sin(x^2) - \sin^2(x)}{x^4} & \quad \text{(m)} \quad \lim_{x \rightarrow \infty} \left[(x^9 + 5x^6)^{1/3} - x^3 \right] \\
 \text{(g)} \quad \lim_{x \rightarrow 0} \frac{e^{-(1/x^2)}}{x} &
 \end{aligned}$$

4. Suppose you want to have \$2000 saved at the end of 5 years. How much money should you place into an account which pays 7% per year compounded continuously?
5. Using a good calculator, find $e^{-0.06} - \left(1 + \frac{0.06}{360}\right)^{360}$. Explain why this gives a measure of the difference between compounding continuously and compounding daily.
6. You know $\lim_{x \rightarrow \infty} \ln x = \infty$. Show that if $\alpha > 0$, then $\lim_{x \rightarrow \infty} \frac{\ln x}{x^\alpha} = 0$.
7. Consider the following function ⁴

$$f(x) = \begin{cases} e^{-1/x^2} & \text{for } x \neq 0 \\ 0 & \text{for } x = 0 \end{cases}$$

Show that $f^{(k)}(0) = 0$ for all k so the power series approximations for this function are all of the form $\sum_{k=0}^m 0x^k$ but the function is not identically equal to 0 on any interval containing 0. Thus this function has all derivatives at 0 and at every other point, yet fails to be approximated by finite sums of the form $\sum_{k=0}^m \frac{f^{(k)}(0)}{k!} x^k$. This is an example of a smooth function which is not analytic. (Roughly speaking, a function is analytic when its power series just described approximates the function near the point at which all the derivatives are evaluated.) It is smooth because all

⁴Surprisingly, this function is very important to those who use modern techniques to study differential equations. One needs to consider test functions which have the property they have infinitely many derivatives but vanish outside of some interval. The theory of complex variables can be used to show there are no examples of such functions if they have a valid power series expansion. It even becomes a little questionable whether such strange functions even exist at all. Nevertheless, they do, there are enough of them, and it is this very example which is used to show this.

derivatives exist and are continuous. It fails to be analytic because $\sum_{k=0}^m \frac{f^{(k)}(0)}{k!} x^k$ fails to approximate the function at any nonzero point because it always gives 0 no matter how large an m is chosen and yet $e^{-1/x^2} \neq 0$ if $x \neq 0$. In fact, there is a sequence of polynomials which will approximate this function on an interval $[0, 1]$, but they are not obtained in the way just described as partial sums of a power series. See Section [4.10](#).

5.20 Videos

[Derivative of Inverse antiderivatives and integrals](#)

Chapter 6

Infinite Series

6.1 Basic Considerations

Earlier in Definition 3.3.1 on Page 86 the notion of limit of a sequence was discussed. There is a very closely related concept called an infinite series which is dealt with in this section.

Definition 6.1.1 Define $\sum_{k=m}^{\infty} a_k \equiv \lim_{n \rightarrow \infty} \sum_{k=m}^n a_k$ whenever the limit exists and is finite. In this case the series is said to converge. If it does not converge, it is said to diverge. The sequence $\{\sum_{k=m}^n a_k\}_{n=m}^{\infty}$ in the above is called the sequence of partial sums. This is always the definition. Here it is understood that the a_k are in \mathbb{R} , but it is the same definition in any situation.

From this definition, it should be clear that infinite sums do not always make sense. Sometimes they do and sometimes they don't, depending on the behavior of the partial sums. As an example, consider $\sum_{k=1}^{\infty} (-1)^k$. The partial sums corresponding to this symbol alternate between -1 and 0 . Therefore, there is no limit for the sequence of partial sums. It follows the symbol just written is meaningless and the infinite sum diverges.

Example 6.1.2 Find the infinite sum, $\sum_{n=1}^{\infty} \frac{1}{n(n+1)}$.

Note $\frac{1}{n(n+1)} = \frac{1}{n} - \frac{1}{n+1}$ and so $\sum_{n=1}^N \frac{1}{n(n+1)} = \sum_{n=1}^N \left(\frac{1}{n} - \frac{1}{n+1} \right) = -\frac{1}{N+1} + 1$. Therefore,

$$\lim_{N \rightarrow \infty} \sum_{n=1}^N \frac{1}{n(n+1)} = \lim_{N \rightarrow \infty} \left(-\frac{1}{N+1} + 1 \right) = 1.$$

Lemma 6.1.3 If $\{A_n\}$ is an increasing sequence in $[-\infty, \infty]$, then $\sup \{A_n\} = \lim_{n \rightarrow \infty} A_n$. If $\{A_n\}$ is a decreasing sequence, then $\inf \{A_n\} = \lim_{n \rightarrow \infty} A_n$.

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 $\lim_{n \rightarrow \infty} A_n = r$ 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 $\lim_{n \rightarrow \infty} A_n = \infty$. The other case is that $r = -\infty$. But in this case, $A_n = -\infty$ for all n and so $\lim_{n \rightarrow \infty} A_n = -\infty$. The other claim is shown the same way. ■

Proposition 6.1.4 *Let $a_k \geq 0$. Then $\{\sum_{k=m}^n a_k\}_{n=m}^\infty$ is an increasing sequence. If this sequence is bounded above, then $\sum_{k=m}^\infty a_k$ converges and its value equals*

$$\sup \left\{ \sum_{k=m}^n a_k : n = m, m+1, \dots \right\}.$$

When the sequence is not bounded above, $\sum_{k=m}^\infty a_k$ diverges. However, in this case, people sometimes write $\sum_{k=m}^\infty a_k = \infty$.

Proof: It follows $\{\sum_{k=m}^n a_k\}_{n=m}^\infty$ is an increasing sequence because $\sum_{k=m}^{n+1} a_k - \sum_{k=m}^n a_k = a_{n+1} \geq 0$. If the sequence of partial sums is bounded above, then this sequence of partial sums must converge to $S \equiv \sup \{\sum_{k=m}^n a_k : n \geq m\}$ by Lemma 6.1.3. If the sequence of partial sums is not bounded, then it cannot converge because if it converged to S , then for all n large enough, $|\sum_{k=m}^n a_k - S| < 1$, and for all such n , $\sum_{k=m}^n a_k \in (1-S, 1+S)$, and there are only finitely many other terms so $\{\sum_{k=m}^n a_k\}$ would need to be bounded. ■

In the case where $a_k \geq 0$, the above proposition shows there are only two alternatives available. Either the sequence of partial sums is bounded above or it is not bounded above. In the first case convergence occurs and in the second case, the infinite series diverges. For this reason, people will sometimes write $\sum_{k=m}^\infty a_k < \infty$ to denote the case where convergence occurs and $\sum_{k=m}^\infty a_k = \infty$ for the case where divergence occurs. Be very careful you never think this way in the case where it is not true that all $a_k \geq 0$. For example, the partial sums of $\sum_{k=1}^\infty (-1)^k$ are bounded because they are all either -1 or 0 but the series does not converge.

One of the most important examples of a convergent series is the geometric series. This series is $\sum_{n=0}^\infty r^n$. The study of this series depends on simple high school algebra and Theorem 3.3.10 on Page 89. Let $S_n \equiv \sum_{k=0}^n r^k$. Then $S_n = \sum_{k=0}^n r^k$, $rS_n = \sum_{k=0}^n r^{k+1} = \sum_{k=1}^{n+1} r^k$. Therefore, subtracting the second equation from the first yields $(1-r)S_n = 1 - r^{n+1}$ and so a formula for S_n is available. In fact, if $r \neq 1$, $S_n = \frac{1-r^{n+1}}{1-r}$. By Theorem 3.3.10, $\lim_{n \rightarrow \infty} S_n = \frac{1}{1-r}$ in the case when $|r| < 1$. Now if $|r| \geq 1$, the limit clearly does not exist because S_n fails to be a Cauchy sequence (Why?) so by Theorem 3.7.3 it cannot converge. This shows the following.

Theorem 6.1.5 *The geometric series, $\sum_{n=0}^\infty r^n$ converges and equals $\frac{1}{1-r}$ if $|r| < 1$ and diverges if $|r| \geq 1$.*

If the series do converge, the following holds.

Theorem 6.1.6 *If $\sum_{k=m}^\infty a_k$ and $\sum_{k=m}^\infty b_k$ both converge and x, y are numbers, then*

$$\sum_{k=m}^\infty a_k = \sum_{k=m+j}^\infty a_{k-j} \quad (6.1)$$

$$\sum_{k=m}^\infty xa_k + yb_k = x \sum_{k=m}^\infty a_k + y \sum_{k=m}^\infty b_k \quad (6.2)$$

$$\left| \sum_{k=m}^\infty a_k \right| \leq \sum_{k=m}^\infty |a_k| \quad (6.3)$$

where in the last inequality, the last sum equals $+\infty$ if the partial sums are not bounded above.

Proof: The above theorem is really only a restatement of Theorem 3.3.7 on Page 87 and the above definitions of infinite series. Thus

$$\sum_{k=m}^{\infty} a_k = \lim_{n \rightarrow \infty} \sum_{k=m}^n a_k = \lim_{n \rightarrow \infty} \sum_{k=m+j}^{n+j} a_{k-j} = \sum_{k=m+j}^{\infty} a_{k-j}.$$

To establish 6.2, use Theorem 3.3.7 on Page 87 to write

$$\begin{aligned} \sum_{k=m}^{\infty} xa_k + yb_k &= \lim_{n \rightarrow \infty} \sum_{k=m}^n xa_k + yb_k = \lim_{n \rightarrow \infty} \left(x \sum_{k=m}^n a_k + y \sum_{k=m}^n b_k \right) \\ &= x \sum_{k=m}^{\infty} a_k + y \sum_{k=m}^{\infty} b_k. \end{aligned}$$

Formula 6.3 follows from the observation that, from the triangle inequality,

$$\left| \sum_{k=m}^n a_k \right| \leq \sum_{k=m}^n |a_k|$$

and so $|\sum_{k=m}^{\infty} a_k| = \lim_{n \rightarrow \infty} |\sum_{k=m}^n a_k| \leq \sum_{k=m}^{\infty} |a_k|$. ■

Recall that if $\lim_{n \rightarrow \infty} A_n = A$, then $\lim_{n \rightarrow \infty} |A_n| = |A|$.

Example 6.1.7 Find $\sum_{n=0}^{\infty} \left(\frac{5}{2^n} + \frac{6}{3^n} \right)$.

From the above theorem and Theorem 6.1.5, $\sum_{n=0}^{\infty} \left(\frac{5}{2^n} + \frac{6}{3^n} \right) =$

$$5 \sum_{n=0}^{\infty} \frac{1}{2^n} + 6 \sum_{n=0}^{\infty} \frac{1}{3^n} = 5 \frac{1}{1 - (1/2)} + 6 \frac{1}{1 - (1/3)} = 19.$$

The following criterion is useful in checking convergence. All it is saying is that the series converges if and only if the sequence of partial sums is Cauchy. This is what the given criterion says. It is just a re-statement of Theorem 3.7.3 on Page 98. It is not new information.

Theorem 6.1.8 Let $\{a_k\}$ be a sequence of points in \mathbb{R} . The sum $\sum_{k=m}^{\infty} a_k$ converges if and only if for all $\varepsilon > 0$, there exists n_ε such that if $q \geq p \geq n_\varepsilon$, then

$$\left| \sum_{k=p}^q a_k \right| < \varepsilon. \quad (6.4)$$

Proof: Suppose first that the series converges. Then $\{\sum_{k=m}^n a_k\}_{n=m}^{\infty}$ is a Cauchy sequence by Theorem 3.7.3 on Page 98. Therefore, there exists $n_\varepsilon > m$ such that if $q \geq p - 1 \geq n_\varepsilon > m$,

$$\left| \sum_{k=m}^q a_k - \sum_{k=m}^{p-1} a_k \right| = \left| \sum_{k=p}^q a_k \right| < \varepsilon. \quad (6.5)$$

Next suppose 6.4 holds. Then from 6.5 it follows upon letting p be replaced with $p + 1$ that $\{\sum_{k=m}^n a_k\}_{n=m}^{\infty}$ is a Cauchy sequence and so, by Theorem 3.7.3, it converges. By the definition of infinite series, this shows the infinite sum converges as claimed. ■

6.2 Absolute Convergence

Absolute convergence is the best kind. It says that if you replace each term with its absolute value, the resulting series converges.

Definition 6.2.1 A series $\sum_{k=m}^{\infty} a_k$ is converges absolutely if $\sum_{k=m}^{\infty} |a_k|$ converges. If the series does converge but does not converge absolutely, then it is said to converge conditionally.

Theorem 6.2.2 If $\sum_{k=m}^{\infty} a_k$ converges absolutely, then it converges.

Proof: Let $\varepsilon > 0$ be given. Then by assumption and Theorem 6.1.8, there exists n_ε such that whenever $q \geq p \geq n_\varepsilon$, $\sum_{k=p}^q |a_k| < \varepsilon$. Therefore, from the triangle inequality, $\varepsilon > \sum_{k=p}^q |a_k| \geq \left| \sum_{k=p}^q a_k \right|$. By Theorem 6.1.8, $\sum_{k=m}^{\infty} a_k$ converges. ■

In fact, the above theorem is really another version of the completeness axiom. Thus its validity implies completeness. You might try to show this.

One of the interesting things about absolutely convergent series is that you can “add them up” in any order and you will always get the same thing. This is the meaning of the following theorem. Of course there is no problem when you are dealing with finite sums thanks to the commutative law of addition. However, when you have infinite sums strange and wonderful things can happen because these involve a limit.

Theorem 6.2.3 Let $\theta : \mathbb{N} \rightarrow \mathbb{N}$ be one to one and onto. Suppose $\sum_{k=1}^{\infty} a_k$ converges absolutely. Then $\sum_{k=1}^{\infty} a_{\theta(k)} = \sum_{k=1}^{\infty} a_k$.

Proof: From absolute convergence, there exists M such that $\sum_{k=M+1}^{\infty} |a_k| < \varepsilon$. Since θ is one to one and onto, there exists $N \geq M$ such that $\{1, 2, \dots, M\} \subseteq \{\theta(1), \theta(2), \dots, \theta(N)\}$. It follows that it is also the case that $\sum_{k=N+1}^{\infty} |a_{\theta(k)}| < \varepsilon$. This is because the partial sums of the above series are each dominated by a partial sum for $\sum_{k=M+1}^{\infty} |a_k|$ since every index $\theta(k)$ equals some n for $n \geq M+1$. Then since ε is arbitrary, this shows that the partial sums of $\sum a_{\theta(k)}$ are Cauchy. Hence, this series does converge and also

$$\left| \sum_{k=1}^M a_k - \sum_{k=1}^N a_{\theta(k)} \right| \leq \sum_{k=M+1}^{\infty} |a_k| < \varepsilon$$

Hence

$$\begin{aligned} \left| \sum_{k=1}^{\infty} a_k - \sum_{k=1}^{\infty} a_{\theta(k)} \right| &\leq \left| \sum_{k=1}^{\infty} a_k - \sum_{k=1}^M a_k \right| + \left| \sum_{k=1}^M a_k - \sum_{k=1}^N a_{\theta(k)} \right| \\ &\quad + \left| \sum_{k=1}^N a_{\theta(k)} - \sum_{k=1}^{\infty} a_{\theta(k)} \right| < \sum_{k=M+1}^{\infty} |a_k| + \varepsilon + \sum_{k=N+1}^{\infty} |a_{\theta(k)}| < 3\varepsilon \end{aligned}$$

Since ε is arbitrary, this shows the two series are equal as claimed. ■

So what happens when series converge only conditionally?

Example 6.2.4 Consider the series $\sum_{k=1}^{\infty} (-1)^k \frac{1}{k}$. Show that there is a rearrangement which converges to 7 although this series does converge. (In fact, it converges to $-\ln 2$.)

First of all consider why it converges. Notice that if S_n denotes the n^{th} partial sum, then

$$\begin{aligned} S_{2n} - S_{2n-2} &= \frac{1}{2n} - \frac{1}{2n-1} < 0 \\ S_{2n+1} - S_{2n-1} &= -\frac{1}{2n+1} + \frac{1}{2n} > 0 \\ S_{2n} - S_{2n-1} &= \frac{1}{2n} \end{aligned}$$

Thus the even partial sums are decreasing and the odd partial sums are increasing. The even partial sums are bounded below also. (Why?) Therefore, the limit of the even partial sums exists. However, it must be the same as the limit of the odd partial sums because of the last equality above. Thus $\lim_{n \rightarrow \infty} S_n$ exists and so the series converges. Now I will show later below that $\sum_k \frac{1}{2k}$ and $\sum_k \frac{1}{2k-1}$ both diverge. Include enough even terms for the sum to exceed 7. Next add in enough odd terms so that the result will be less than 7. Next add enough even terms to exceed 7 and continue doing this. Since $1/k$ converges to 0, this rearrangement of the series must converge to 7. Of course you could also have picked 5 or -8 just as well. In fact, given any number, there is a rearrangement of this series which converges to this number. Calculus is not algebra! No such thing happens with finite sums!

Theorem 6.2.5 (comparison test) Suppose $\{a_n\}$ and $\{b_n\}$ are sequences of non negative real numbers and suppose for all n sufficiently large, $a_n \leq b_n$. Then

1. If $\sum_{n=k}^{\infty} b_n$ converges, then $\sum_{n=m}^{\infty} a_n$ converges.
2. If $\sum_{n=k}^{\infty} a_n$ diverges, then $\sum_{n=m}^{\infty} b_n$ diverges.

Proof: Consider the first claim. From the assumption, there exists n^* such that $n^* > \max(k, m)$ and for all $n \geq n^*$ $b_n \geq a_n$. Then if $p \geq n^*$,

$$\sum_{n=m}^p a_n \leq \sum_{n=m}^{n^*} a_n + \sum_{n=n^*+1}^p b_n \leq \sum_{n=m}^{n^*} a_n + \sum_{n=k}^{\infty} b_n.$$

Thus the sequence, $\{\sum_{n=m}^p a_n\}_{p=m}^{\infty}$ is bounded above and increasing. Therefore, it converges by completeness. The second claim is left as an exercise. ■

Example 6.2.6 Determine the convergence of $\sum_{n=1}^{\infty} \frac{1}{n^2}$.

For $n > 1$, $\frac{1}{n^2} \leq \frac{1}{n(n-1)}$. Now

$$\sum_{n=2}^p \frac{1}{n(n-1)} = \sum_{n=2}^p \left[\frac{1}{n-1} - \frac{1}{n} \right] = 1 - \frac{1}{p} \rightarrow 1 \text{ as } p \rightarrow \infty$$

Therefore, letting $a_n = \frac{1}{n^2}$ and $b_n = \frac{1}{n(n-1)}$ the conclusion follows from Theorem 6.2.5.

A convenient way to implement the comparison test is to use the limit comparison test. This is considered next.

Theorem 6.2.7 Let $a_n, b_n > 0$ and suppose for all n large enough,

$$0 < a < \frac{a_n}{b_n} \leq \frac{a_n}{b_n} < b < \infty.$$

Then $\sum a_n$ and $\sum b_n$ converge or diverge together.

Proof: Let n^* be such that $n \geq n^*$, then $\frac{a_n}{b_n} > a$ and $\frac{a_n}{b_n} < b$ and so for all such n , $ab_n < a_n < bb_n$ and so the conclusion follows from the comparison test. ■

The following corollary follows right away from the definition of the limit.

Corollary 6.2.8 Let $a_n, b_n > 0$ and suppose $\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = \lambda \in (0, \infty)$. Then $\sum a_n$ and $\sum b_n$ converge or diverge together.

Example 6.2.9 Determine the convergence of $\sum_{k=1}^{\infty} \frac{1}{\sqrt{n^4+2n+7}}$.

This series converges by the limit comparison test above. Compare with the series of Example 6.2.6.

$$\lim_{n \rightarrow \infty} \frac{\left(\frac{1}{n^2}\right)}{\left(\frac{1}{\sqrt{n^4+2n+7}}\right)} = \lim_{n \rightarrow \infty} \frac{\sqrt{n^4+2n+7}}{n^2} = \lim_{n \rightarrow \infty} \sqrt{1 + \frac{2}{n^3} + \frac{7}{n^4}} = 1.$$

Therefore, the series converges with the series of Example 6.2.6. How did I know what to compare with? I noticed that $\sqrt{n^4+2n+7}$ is essentially like $\sqrt{n^4} = n^2$ for large enough n . You see, the higher order term n^4 dominates the other terms in n^4+2n+7 . Therefore, reasoning that $1/\sqrt{n^4+2n+7}$ is a lot like $1/n^2$ for large n , it was easy to see what to compare with. Of course this is not always easy and there is room for acquiring skill through practice.

To really exploit this limit comparison test, it is desirable to get lots of examples of series, some which converge and some which do not. The tool for obtaining these examples here will be the following wonderful theorem known as the Cauchy condensation test.

Theorem 6.2.10 Let $a_n \geq 0$ and suppose the terms of the sequence $\{a_n\}$ are decreasing. Thus $a_n \geq a_{n+1}$ for all n . Then $\sum_{n=1}^{\infty} a_n$ and $\sum_{n=0}^{\infty} 2^n a_{2^n}$ converge or diverge together.

Proof: This follows from the inequality of the following claim.

Claim: $\sum_{k=1}^n 2^k a_{2^{k-1}} \geq \sum_{k=1}^{2^n} a_k \geq \sum_{k=0}^{2^n-1} 2^{k-1} a_{2^k}$.

Proof of the Claim: Note the claim is true for $n = 1$. Suppose the claim is true for n . Then, since $2^{n+1} - 2^n = 2^n$, and the terms, a_n , are decreasing,

$$\begin{aligned} \sum_{k=1}^{n+1} 2^k a_{2^{k-1}} &= 2^{n+1} a_{2^n} + \sum_{k=1}^n 2^k a_{2^{k-1}} \geq 2^{n+1} a_{2^n} + \sum_{k=1}^{2^n} a_k \\ &\geq \sum_{k=1}^{2^{n+1}} a_k \geq 2^n a_{2^{n+1}} + \sum_{k=1}^{2^n} a_k \geq 2^n a_{2^{n+1}} + \sum_{k=0}^n 2^{k-1} a_{2^k} = \sum_{k=0}^{n+1} 2^{k-1} a_{2^k}. \quad \blacksquare \end{aligned}$$

In case it is not clear why the claim implies the assertion, consider the case where $\sum_{n=0}^{\infty} 2^n a_{2^n}$ converges. Then $2 \sum_{n=0}^{\infty} 2^n a_{2^n} = \sum_{n=1}^{\infty} 2^n a_{2^{n-1}}$ is finite. Then from the claim, $\sum_{k=1}^{2^n} a_k \leq \sum_{k=1}^{n+1} 2^k a_{2^{k-1}} \leq \sum_{n=1}^{\infty} 2^n a_{2^{n-1}} < \infty$ and so the partial sums are bounded. Since the terms of the series are nonnegative, the infinite series converges as shown earlier. In case $\sum_{n=0}^{\infty} 2^n a_{2^n}$ diverges, a similar argument shows the partial sums of the original series are unbounded.

Example 6.2.11 Determine the convergence of $\sum_{k=1}^{\infty} \frac{1}{k^p}$ where p is a positive number. These are called the p series.

Let $a_n = \frac{1}{n^p}$. Then $a_{2^n} = \left(\frac{1}{2^p}\right)^n$. From the Cauchy condensation test the two series

$$\sum_{n=1}^{\infty} \frac{1}{n^p} \text{ and } \sum_{n=0}^{\infty} 2^n \left(\frac{1}{2^p}\right)^n = \sum_{n=0}^{\infty} \left(2^{(1-p)}\right)^n$$

converge or diverge together. If $p > 1$, the last series above is a geometric series having common ratio less than 1 and so it converges. If $p \leq 1$, it is still a geometric series but in this case the common ratio is either 1 or greater than 1 so the series diverges. It follows that the p series converges if $p > 1$ and diverges if $p \leq 1$. In particular, $\sum_{n=1}^{\infty} n^{-1}$ diverges while $\sum_{n=1}^{\infty} n^{-2}$ converges.

Example 6.2.12 Determine the convergence of $\sum_{k=1}^{\infty} \frac{1}{\sqrt{n^2+100n}}$.

Use the limit comparison test. $\lim_{n \rightarrow \infty} \frac{\left(\frac{1}{n}\right)}{\left(\frac{1}{\sqrt{n^2+100n}}\right)} = 1$ and so this series diverges with

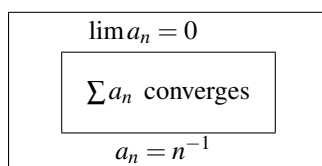
$$\sum_{k=1}^{\infty} \frac{1}{k}.$$

Sometimes it is good to be able to say a series does not converge. The n^{th} term test gives such a condition which is sufficient for this. It is really a corollary of Theorem 6.1.8. Here is the n^{th} term test.

Theorem 6.2.13 If $\sum_{n=m}^{\infty} a_n$ converges, then $\lim_{n \rightarrow \infty} a_n = 0$.

Proof: Apply Theorem 6.1.8 to conclude that $\lim_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} \sum_{k=n}^{\infty} a_k = 0$. ■

It is very important to observe that this theorem goes only in one direction. That is, you cannot conclude the series converges if $\lim_{n \rightarrow \infty} a_n = 0$. If this happens, you don't know anything from this information. Recall $\lim_{n \rightarrow \infty} n^{-1} = 0$ but $\sum_{n=1}^{\infty} n^{-1}$ diverges. The following picture is descriptive of the situation.



6.3 Ratio and Root Tests

A favorite test for convergence is the ratio test. This is discussed next. There are exactly three possible outcomes for this test: failure, spectacular divergence, and absolute convergence.

Theorem 6.3.1 Suppose $|a_n| > 0$ for all n and suppose $\lim_{n \rightarrow \infty} \frac{|a_{n+1}|}{|a_n|} = r$. Then

$$\sum_{n=1}^{\infty} a_n \begin{cases} \text{diverges if } r > 1 \\ \text{converges absolutely if } r < 1 \\ \text{test fails if } r = 1 \end{cases}.$$

Proof: Suppose $r < 1$. Then there exists n_1 such that if $n \geq n_1$, then $0 < \left| \frac{a_{n+1}}{a_n} \right| < R$ where $r < R < 1$. Then $|a_{n+1}| < R|a_n|$ for all such n . Therefore,

$$|a_{n_1+p}| < R|a_{n_1+p-1}| < R^2|a_{n_1+p-2}| < \cdots < R^p|a_{n_1}| \quad (6.6)$$

and so if $m > n$, then $|a_m| < R^{m-n_1}|a_{n_1}|$. By the comparison test and the theorem on geometric series, $\sum |a_n|$ converges. This proves the convergence part of the theorem.

To verify the divergence part, note that if $r > 1$, then 6.6 can be turned around for some $R > 1$. Showing $\lim_{n \rightarrow \infty} |a_n| = \infty$. Since the n^{th} term fails to converge to 0, it follows the series diverges.

To see the test fails if $r = 1$, consider $\sum n^{-1}$ and $\sum n^{-2}$. The first series diverges while the second one converges but in both cases, $r = 1$. (Be sure to check this last claim.) ■

The ratio test is very useful for many different examples but it is somewhat unsatisfactory mathematically. One reason for this is the assumption that $a_n \neq 0$, necessitated by the need to divide by a_n , and the other reason is the possibility that the limit might not exist. The next test, called the root test removes both of these objections.

Theorem 6.3.2 Suppose $|a_n|^{1/n} < R < 1$ for all n sufficiently large. Then $\sum_{n=1}^{\infty} a_n$ converges absolutely. If there are infinitely many values of n such that $|a_n|^{1/n} \geq 1$, then $\sum_{n=1}^{\infty} a_n$ diverges.

Proof: Suppose first that $|a_n|^{1/n} < R < 1$ for all n sufficiently large. Say this holds for all $n \geq n_R$. Then for such n , $\sqrt[n]{|a_n|} < R$. Therefore, for such n , $|a_n| \leq R^n$ and so the comparison test with a geometric series applies and gives absolute convergence as claimed.

Next suppose $|a_n|^{1/n} \geq 1$ for infinitely many values of n . Then for those values of n , $|a_n| \geq 1$ and so the series fails to converge by the n^{th} term test, Theorem 6.2.13. ■

Stated more succinctly, using Definition 3.3.16 the condition for the root test is this: Let $r \equiv \limsup_{n \rightarrow \infty} |a_n|^{1/n}$ then

$$\sum_{k=m}^{\infty} a_k \begin{cases} \text{converges absolutely if } r < 1 \\ \text{test fails if } r = 1 \\ \text{diverges if } r > 1 \end{cases}$$

To see the test fails when $r = 1$, consider the same example given above, $\sum_n \frac{1}{n}$ and $\sum_n \frac{1}{n^2}$.

A special case occurs when the limit exists.

Corollary 6.3.3 Suppose $\lim_{n \rightarrow \infty} |a_n|^{1/n}$ exists and equals r . Then

$$\sum_{k=m}^{\infty} a_k \begin{cases} \text{converges absolutely if } r < 1 \\ \text{test fails if } r = 1 \\ \text{diverges if } r > 1 \end{cases}$$

Proof: The first and last alternatives follow from Theorem 6.3.2. To see the test fails if $r = 1$, consider the two series $\sum_{n=1}^{\infty} \frac{1}{n}$ and $\sum_{n=1}^{\infty} \frac{1}{n^2}$ both of which have $r = 1$ but having different convergence properties. The first diverges and the second converges. ■

6.4 Exercises

1. Determine whether the following series converge and give reasons for your answers.

$$\begin{array}{ll} \text{(a)} \sum_{n=1}^{\infty} \frac{1}{\sqrt{n^2+n+1}} & \text{(e)} \sum_{n=1}^{\infty} \frac{1}{2n+2} \\ \text{(b)} \sum_{n=1}^{\infty} (\sqrt{n+1} - \sqrt{n}) & \text{(f)} \sum_{n=1}^{\infty} \left(\frac{n}{n+1}\right)^n \\ \text{(c)} \sum_{n=1}^{\infty} \frac{(n!)^2}{(2n)!} & \text{(g)} \sum_{n=1}^{\infty} \left(\frac{n}{n+1}\right)^{n^2} \\ \text{(d)} \sum_{n=1}^{\infty} \frac{(2n)!}{(n!)^2} \end{array}$$

2. Determine whether the following series converge and give reasons for your answers.

$$\begin{array}{ll} \text{(a)} \sum_{n=1}^{\infty} \frac{2^n+n}{n2^n} & \text{(c)} \sum_{n=1}^{\infty} \frac{n}{2n+1} \\ \text{(b)} \sum_{n=1}^{\infty} \frac{2^n+n}{n^2 2^n} & \text{(d)} \sum_{n=1}^{\infty} \frac{n^{100}}{1.01^n} \end{array}$$

3. Find the exact values of the following infinite series if they converge.

$$\begin{array}{ll} \text{(a)} \sum_{k=3}^{\infty} \frac{1}{k(k-2)} & \text{(c)} \sum_{k=3}^{\infty} \frac{1}{(k+1)(k-2)} \\ \text{(b)} \sum_{k=1}^{\infty} \frac{1}{k(k+1)} & \text{(d)} \sum_{k=1}^{\infty} \left(\frac{1}{\sqrt{k}} - \frac{1}{\sqrt{k+1}}\right) \end{array}$$

4. Suppose $\sum_{k=1}^{\infty} a_k$ converges and each $a_k \geq 0$. Does it follow that $\sum_{k=1}^{\infty} a_k^2$ also converges?

5. Find a series which diverges using one test but converges using another if possible. If this is not possible, tell why.

6. If $\sum_{n=1}^{\infty} a_n$ and $\sum_{n=1}^{\infty} b_n$ both converge and a_n, b_n are nonnegative, can you conclude the sum, $\sum_{n=1}^{\infty} a_n b_n$ converges?

7. If $\sum_{n=1}^{\infty} a_n$ converges and $a_n \geq 0$ for all n and b_n is bounded, can you conclude $\sum_{n=1}^{\infty} a_n b_n$ converges?

8. Determine the convergence of the series $\sum_{n=1}^{\infty} \left(\sum_{k=1}^n \frac{1}{k}\right)^{-n/2}$.

9. Is it possible there could exist a decreasing sequence of positive numbers, $\{a_n\}$ such that $\lim_{n \rightarrow \infty} a_n = 0$ but $\sum_{n=1}^{\infty} \left(1 - \frac{a_{n+1}}{a_n}\right)$ converges? (This seems to be a fairly difficult problem.) **Hint:** You might do something like this. Show $\lim_{x \rightarrow 1} \frac{1-x}{-\ln(x)} = \frac{1-x}{\ln(1/x)} = 1$. Next use a limit comparison test with $\sum_{n=1}^{\infty} \ln\left(\frac{a_n}{a_{n+1}}\right)$.

10. Suppose $\sum a_n$ converges conditionally and each a_n is real. Show it is possible to add the series in some order such that the result converges to 13. Then show it is possible to add the series in another order so that the result converges to 7. Thus there is no generalization of the commutative law for conditionally convergent infinite series.

Hint: To see how to proceed, consider Example 6.2.4.

11. He takes a drug every evening and after 8 hours there is half of it left. Find upper and lower bounds for the amount of drug in his body if he has been taking it for a long time. Assume each dose consists of 10 mg.

6.5 Convergence Because of Cancellation

So far, the tests for convergence have been applied to non negative terms only. Sometimes, a series converges, not because the terms of the series get small fast enough, but because of cancellation taking place between positive and negative terms. A discussion of this involves some simple algebra and yields a much more subtle test for convergence.

Let $\{a_n\}$ and $\{b_n\}$ be sequences and let $A_n \equiv \sum_{k=1}^n a_k$, $A_{-1} \equiv A_0 \equiv 0$. Then if $p < q$

$$\begin{aligned} \sum_{n=p}^q a_n b_n &= \sum_{n=p}^q b_n (A_n - A_{n-1}) = \sum_{n=p}^q b_n A_n - \sum_{n=p}^q b_n A_{n-1} \\ &= \sum_{n=p}^q b_n A_n - \sum_{n=p-1}^{q-1} b_{n+1} A_n = b_q A_q - b_p A_{p-1} + \sum_{n=p}^{q-1} A_n (b_n - b_{n+1}) \end{aligned} \quad (6.7)$$

This formula is called the partial summation formula. It is just like integration by parts. This yields Dirichlet's test ¹.

Theorem 6.5.1 (*Dirichlet's test*) Suppose $A_n \equiv \sum_{k=1}^n a_k$ is bounded independent of n , meaning for some $C > 0$, $|A_n| \leq C$ for all n . and $\lim_{n \rightarrow \infty} b_n = 0$, with $b_n \geq b_{n+1}$ for all n . Then $\sum a_n b_n$ converges. Thus it makes perfect sense to write $\sum_{n=1}^{\infty} a_n b_n$.

Proof: This follows quickly from Theorem 6.1.8. Indeed, letting $|A_n| \leq C$, and using the partial summation formula above along with the assumption that the b_n are decreasing,

$$\begin{aligned} \left| \sum_{n=p}^q a_n b_n \right| &= \left| b_q A_q - b_p A_{p-1} + \sum_{n=p}^{q-1} A_n (b_n - b_{n+1}) \right| \\ &\leq C (|b_q| + |b_p|) + C \sum_{n=p}^{q-1} (b_n - b_{n+1}) = C (|b_q| + |b_p|) + C (b_p - b_q) \end{aligned}$$

and by assumption, this last expression is small whenever p and q are sufficiently large. Thus the partial sums are a Cauchy sequence. ■

Definition 6.5.2 If $b_n > 0$ for all n , a series of the form $\sum_k (-1)^k b_k$ or $\sum_k (-1)^{k-1} b_k$ is known as an alternating series.

The following corollary is known as the alternating series test.

Corollary 6.5.3 (*alternating series test*) If $\lim_{n \rightarrow \infty} b_n = 0$, with $b_n \geq b_{n+1}$, then it follows that the series $\sum_{n=1}^{\infty} (-1)^n b_n$ converges.

Proof: Let $a_n = (-1)^n$. Then the partial sums of $\sum_n a_n$ are bounded and so Theorem 6.5.1 applies. ■

In the situation of Corollary 6.5.3 there is a convenient error estimate available.

¹Peter Gustav Lejeune Dirichlet, 1805-1859 was a German mathematician who did fundamental work in analytic number theory. He also gave the first proof that Fourier series tend to converge to the mid-point of the jump of the function. He is a very important figure in the development of analysis in the nineteenth century. An interesting personal fact is that the great composer Felix Mendelssohn was his brother in law.

Theorem 6.5.4 Let $b_n > 0$ for all n such that $b_n \geq b_{n+1}$ for all n and $\lim_{n \rightarrow \infty} b_n = 0$ and consider either $\sum_{n=1}^{\infty} (-1)^n b_n$ or $\sum_{n=1}^{\infty} (-1)^{n-1} b_n$. Then

$$\left| \sum_{n=1}^{\infty} (-1)^n b_n - \sum_{n=1}^N (-1)^n b_n \right| \leq |b_{N+1}|,$$

$$\left| \sum_{n=1}^{\infty} (-1)^{n-1} b_n - \sum_{n=1}^N (-1)^{n-1} b_n \right| \leq |b_{N+1}|$$

See Problem 8 on Page 177 for an outline of the proof of this theorem along with another way to prove the alternating series test.

Example 6.5.5 How many terms must I take in the sum, $\sum_{n=1}^{\infty} (-1)^n \frac{1}{n^2+1}$ to be closer than $\frac{1}{10}$ to $\sum_{n=1}^{\infty} (-1)^n \frac{1}{n^2+1}$?

From Theorem 6.5.4, I need to find n such that $\frac{1}{n^2+1} \leq \frac{1}{10}$ and then $n-1$ is the desired value. Thus $n = 3$ and so

$$\left| \sum_{n=1}^{\infty} (-1)^n \frac{1}{n^2+1} - \sum_{n=1}^2 (-1)^n \frac{1}{n^2+1} \right| \leq \frac{1}{10}$$

Definition 6.5.6 A series $\sum a_n$ is said to converge absolutely if $\sum |a_n|$ converges. It is said to converge conditionally if $\sum |a_n|$ fails to converge but $\sum a_n$ converges.

Thus the alternating series or more general Dirichlet test can determine convergence of series which converge conditionally.

6.6 Double Series

Sometimes it is required to consider double series which 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. Whenever you interchange the order in which two limits are taken, you need a theorem which will allow you to do it. Such theorems are often rather technical. One must never interchange limits of any kind without agonizing over whether the symbol pushing is correct. In general, limits ruin algebra and also introduce things which are counter intuitive. Failure to keep this in mind leads to mathematical disasters. 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 6.6.1 Consider the following picture which depicts some of the ordered pairs (m, n) where m, n are positive integers.

$$\begin{array}{cccccc}
 & & & & & \vdots \\
 & & & & & 0 & 0 & c & 0 & -c \\
 & & & & & 0 & c & 0 & -c & 0 \\
 & & & & & b & 0 & -c & 0 & 0 & \dots \\
 & & & & & 0 & a & 0 & 0 & 0
 \end{array}$$

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.

Don't become upset by this. It happens because, as indicated above, limits are taken in two different orders. An infinite sum always involves a limit and this illustrates why you must always remember this. This example in no way violates the commutative law of addition which has nothing to do with limits. However, 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 6.6.2 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 6.6.3 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 6.6.4 *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, using Lemma 6.1.3,

$$\begin{aligned} \sum_{j=r}^{\infty} \sum_{i=r}^{\infty} a_{ij} &\geq \sup_n \sum_{j=r}^{\infty} \sum_{i=r}^n a_{ij} = \sup_n \lim_{m \rightarrow \infty} \sum_{j=r}^m \sum_{i=r}^n a_{ij} = \sup_n \lim_{m \rightarrow \infty} \sum_{i=r}^n \sum_{j=r}^m a_{ij} \\ &= \sup_n \sum_{i=r}^n \lim_{m \rightarrow \infty} \sum_{j=r}^m a_{ij} = \sup_n \sum_{i=r}^n \sum_{j=r}^{\infty} a_{ij} = \lim_{n \rightarrow \infty} \sum_{i=r}^n \sum_{j=r}^{\infty} a_{ij} = \sum_{i=r}^{\infty} \sum_{j=r}^{\infty} a_{ij} \end{aligned}$$

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

The following is the fundamental result on double sums.

Theorem 6.6.5 *Let $a_{ij} \in \mathbb{R}$ and suppose $\sum_{i=r}^{\infty} \sum_{j=r}^{\infty} |a_{ij}| < \infty$. Then $\sum_{i=r}^{\infty} \sum_{j=r}^{\infty} a_{ij} = \sum_{j=r}^{\infty} \sum_{i=r}^{\infty} a_{ij}$ and every infinite sum encountered in the above equation converges.*

Proof: By Theorem 6.6.4, $\sum_{j=r}^{\infty} \sum_{i=r}^{\infty} |a_{ij}| = \sum_{i=r}^{\infty} \sum_{j=r}^{\infty} |a_{ij}| < \infty$. Therefore, for each j , $\sum_{i=r}^{\infty} |a_{ij}| < \infty$ and for each i , $\sum_{j=r}^{\infty} |a_{ij}| < \infty$. By Theorem 6.2.2 on Page 166, both of the series $\sum_{i=r}^{\infty} a_{ij}$, $\sum_{j=r}^{\infty} a_{ij}$ converge, the first one for every j and the second for every i . Also, $\sum_{j=r}^{\infty} |\sum_{i=r}^{\infty} a_{ij}| \leq \sum_{j=r}^{\infty} \sum_{i=r}^{\infty} |a_{ij}| < \infty$ and $\sum_{i=r}^{\infty} |\sum_{j=r}^{\infty} a_{ij}| \leq \sum_{i=r}^{\infty} \sum_{j=r}^{\infty} |a_{ij}| < \infty$ so by Theorem 6.2.2 again, $\sum_{j=r}^{\infty} \sum_{i=r}^{\infty} a_{ij}$, $\sum_{i=r}^{\infty} \sum_{j=r}^{\infty} a_{ij}$ both exist. It only remains to verify they are equal.

By Theorem 6.6.4 and Theorem 6.1.6 on Page 164

$$\begin{aligned} \sum_{j=r}^{\infty} \sum_{i=r}^{\infty} |a_{ij}| + \sum_{j=r}^{\infty} \sum_{i=r}^{\infty} a_{ij} &= \sum_{j=r}^{\infty} \sum_{i=r}^{\infty} (|a_{ij}| + a_{ij}) \\ &= \sum_{i=r}^{\infty} \sum_{j=r}^{\infty} (|a_{ij}| + a_{ij}) = \sum_{i=r}^{\infty} \sum_{j=r}^{\infty} |a_{ij}| + \sum_{i=r}^{\infty} \sum_{j=r}^{\infty} a_{ij} = \sum_{j=r}^{\infty} \sum_{i=r}^{\infty} |a_{ij}| + \sum_{j=r}^{\infty} \sum_{i=r}^{\infty} a_{ij} \end{aligned}$$

and so $\sum_{j=r}^{\infty} \sum_{i=r}^{\infty} a_{ij} = \sum_{i=r}^{\infty} \sum_{j=r}^{\infty} a_{ij}$. It follows the two series are equal. ■

One of the most important applications of this theorem is to the problem of multiplication of series.

Definition 6.6.6 *Let $\sum_{i=r}^{\infty} a_i$ and $\sum_{i=r}^{\infty} b_i$ be two series. For $n \geq r$, define*

$$c_n \equiv \sum_{k=r}^n a_k b_{n-k+r}.$$

The series $\sum_{n=r}^{\infty} c_n$ is called the Cauchy product of the two series.

It isn't hard to see where this comes from. Formally write the following in the case $r = 0$:

$$(a_0 + a_1 + a_2 + a_3 \cdots)(b_0 + b_1 + b_2 + b_3 \cdots)$$

and start multiplying in the usual way. This yields

$$a_0b_0 + (a_0b_1 + b_0a_1) + (a_0b_2 + a_1b_1 + a_2b_0) + \cdots$$

and you see the expressions in parentheses above are just the c_n for $n = 0, 1, 2, \dots$. Therefore, it is reasonable to conjecture that $\sum_{i=r}^{\infty} a_i \sum_{j=r}^{\infty} b_j = \sum_{n=r}^{\infty} c_n$ and of course there would be no problem with this in the case of finite sums but in the case of infinite sums, it is necessary to prove a theorem. The following is a special case of Merten's theorem.

Theorem 6.6.7 Suppose $\sum_{i=r}^{\infty} a_i$ and $\sum_{j=r}^{\infty} b_j$ both converge absolutely². Then

$$\left(\sum_{i=r}^{\infty} a_i \right) \left(\sum_{j=r}^{\infty} b_j \right) = \sum_{n=r}^{\infty} c_n$$

where $c_n = \sum_{k=r}^n a_k b_{n-k+r}$.

Proof: Let $p_{nk} = 1$ if $r \leq k \leq n$ and $p_{nk} = 0$ if $k > n$. Then $c_n = \sum_{k=r}^{\infty} p_{nk} a_k b_{n-k+r}$. Also,

$$\begin{aligned} \sum_{k=r}^{\infty} \sum_{n=r}^{\infty} p_{nk} |a_k| |b_{n-k+r}| &= \sum_{k=r}^{\infty} |a_k| \sum_{n=r}^{\infty} p_{nk} |b_{n-k+r}| \\ &= \sum_{k=r}^{\infty} |a_k| \sum_{n=k}^{\infty} |b_{n-k+r}| = \sum_{k=r}^{\infty} |a_k| \sum_{n=k}^{\infty} |b_{n-(k-r)}| = \sum_{k=r}^{\infty} |a_k| \sum_{m=r}^{\infty} |b_m| < \infty. \end{aligned}$$

Therefore, by Theorem 6.6.5

$$\begin{aligned} \sum_{n=r}^{\infty} c_n &= \sum_{n=r}^{\infty} \sum_{k=r}^n a_k b_{n-k+r} = \sum_{n=r}^{\infty} \sum_{k=r}^{\infty} p_{nk} a_k b_{n-k+r} \\ &= \sum_{k=r}^{\infty} a_k \sum_{n=r}^{\infty} p_{nk} b_{n-k+r} = \sum_{k=r}^{\infty} a_k \sum_{n=k}^{\infty} b_{n-k+r} = \sum_{k=r}^{\infty} a_k \sum_{m=r}^{\infty} b_m \quad \blacksquare \end{aligned}$$

6.7 Exercises

1. Determine whether the following series converge absolutely, conditionally, or not at all and give reasons for your answers.

(a) $\sum_{n=1}^{\infty} (-1)^n \frac{2^n + n}{n 2^n}$

(f) $\sum_{n=1}^{\infty} (-1)^n \frac{3^n}{n^3}$

(b) $\sum_{n=1}^{\infty} (-1)^n \frac{2^n + n}{n^2 2^n}$

(g) $\sum_{n=1}^{\infty} (-1)^n \frac{n^3}{3^n}$

(c) $\sum_{n=1}^{\infty} (-1)^n \frac{n}{2n+1}$

(h) $\sum_{n=1}^{\infty} (-1)^n \frac{n^3}{n!}$

(d) $\sum_{n=1}^{\infty} (-1)^n \frac{10^n}{n!}$

(i) $\sum_{n=1}^{\infty} (-1)^n \frac{n!}{n^{100}}$

(e) $\sum_{n=1}^{\infty} (-1)^n \frac{n^{100}}{1.01^n}$

²Actually, it is only necessary to assume one of the series converges and the other converges absolutely. This is known as Merten's theorem and may be read in the 1974 book by Apostol listed in the bibliography.

2. Suppose $\sum_{n=1}^{\infty} a_n$ converges. Can the same thing be said about $\sum_{n=1}^{\infty} a_n^2$? Explain.
3. A person says a series converges conditionally by the ratio test. Explain why his statement is total nonsense.
4. A person says a series diverges by the alternating series test. Explain why his statement is total nonsense.
5. Find a series which diverges using one test but converges using another if possible. If this is not possible, tell why.
6. If $\sum_{n=1}^{\infty} a_n$ and $\sum_{n=1}^{\infty} b_n$ both converge, does $\sum_{n=1}^{\infty} a_n b_n$ converge?
7. If $\sum_{n=1}^{\infty} a_n$ converges absolutely, and b_n is bounded, does $\sum_{n=1}^{\infty} a_n b_n$ converge? What if it is only the case that $\sum_{n=1}^{\infty} a_n$ converges?
8. Prove Theorem 6.5.4. **Hint:** For $\sum_{n=1}^{\infty} (-1)^n b_n$, show the odd partial sums are all no larger than $\sum_{n=1}^{\infty} (-1)^n b_n$ and are increasing while the even partial sums are at least as large as $\sum_{n=1}^{\infty} (-1)^n b_n$ and are decreasing. Use this to give another proof of the alternating series test. If you have trouble, see most standard calculus books.
9. Use Theorem 6.5.4 in the following alternating series to tell how large n must be so that $\left| \sum_{k=1}^{\infty} (-1)^k a_k - \sum_{k=1}^n (-1)^k a_k \right|$ is no larger than the given number.
 - (a) $\sum_{k=1}^{\infty} (-1)^k \frac{1}{k}, .001$
 - (b) $\sum_{k=1}^{\infty} (-1)^k \frac{1}{k^2}, .001$
 - (c) $\sum_{k=1}^{\infty} (-1)^{k-1} \frac{1}{\sqrt{k}}, .001$
10. Consider the series $\sum_{n=0}^{\infty} (-1)^n \frac{1}{\sqrt{n+1}}$. Show this series converges and so it makes sense to write $\left(\sum_{n=0}^{\infty} (-1)^n \frac{1}{\sqrt{n+1}} \right)^2$. What about the Cauchy product of this series? Does it even converge? What does this mean about using algebra on infinite sums as though they were finite sums?
11. Verify Theorem 6.6.7 on the two series $\sum_{k=0}^{\infty} 2^{-k}$ and $\sum_{k=0}^{\infty} 3^{-k}$.
12. All of the above involves only real sums of real numbers. However, you can define infinite series of complex numbers in exactly the same way as infinite series of real numbers. That is $w = \sum_{k=1}^{\infty} z_k$ means: For every $\varepsilon > 0$ there exists N such that if $n \geq N$, then $|w - \sum_{k=1}^n z_k| < \varepsilon$. Here the absolute value is the one which applies to complex numbers. That is, $|a + ib| = \sqrt{a^2 + b^2}$. Show that if $\{a_n\}$ is a decreasing sequence of nonnegative numbers with the property that $\lim_{n \rightarrow \infty} a_n = 0$ and if ω is any complex number which is not equal to 1 but which satisfies $|\omega| = 1$, then $\sum_{n=1}^{\infty} \omega^n a_n$ must converge. Note a sequence of complex numbers, $\{a_n + ib_n\}$ converges to $a + ib$ if and only if $a_n \rightarrow a$ and $b_n \rightarrow b$. There are quite a few things in this problem you should think about.
13. Suppose $\lim_{k \rightarrow \infty} s_k = s$. Show it follows $\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n s_k = s$.
14. Using Problem 13 show that if $\sum_{j=1}^{\infty} \frac{a_j}{j}$ converges, then $\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n a_j = 0$.

15. Show that if $\{p_i\}_{i=1}^{\infty}$ are the prime numbers, then $\sum_{i=1}^{\infty} \frac{1}{p_i} = \infty$. That is, there are enough primes that the sum of their reciprocals diverges. **Hint:** Let $\pi(n)$ denote the number of primes less than equal to n , $\{p_1, \dots, p_{\pi(n)}\}$. Then explain why

$$\sum_{k=1}^n \frac{1}{k} \leq \left(\sum_{k=1}^n \frac{1}{p_1^k} \right) \cdots \left(\sum_{k=1}^n \frac{1}{p_{\pi(n)}^k} \right) \leq \prod_{k=1}^{\pi(n)} \frac{1}{1 - \frac{1}{p_k}} \leq \prod_{k=1}^{\pi(n)} e^{2/p_k} = e^{2 \sum_{k=1}^{\pi(n)} \frac{1}{p_k}}$$

and consequently why $\lim_{n \rightarrow \infty} \pi(n) = \infty$ and $\sum_{i=1}^{\infty} \frac{1}{p_i} = \infty$.

6.8 Series of Functions

Infinite sequences of functions were discussed earlier. Remember, there were two kinds of convergence, pointwise and uniform. As was just done for series of numbers, once you understand sequences, it is no problem to consider series. In this case, series of functions.

Definition 6.8.1 Let $\{f_n\}$ be a sequence of functions defined on D . Then

$$\left(\sum_{k=1}^{\infty} f_k \right) (x) \equiv \lim_{n \rightarrow \infty} \sum_{k=1}^n f_k(x) \quad (6.8)$$

whenever the limit exists. Thus there is a new function denoted by

$$\sum_{k=1}^{\infty} f_k \quad (6.9)$$

and its value at x is given by the limit of the sequence of partial sums in 6.8. If for all $x \in D$, the limit in 6.8 exists, then 6.9 is said to converge pointwise. $\sum_{k=1}^{\infty} f_k$ is said to converge uniformly on D if the sequence of partial sums, $\{\sum_{k=1}^n f_k\}_{n=1}^{\infty}$ converges uniformly. If the indices for the functions start at some other value than 1, you make the obvious modification to the above definition as was done earlier with series of numbers.

Theorem 6.8.2 Let $\{f_n\}$ be a sequence of functions defined on D . The series $\sum_{k=1}^{\infty} f_k$ converges pointwise if and only if for each $\varepsilon > 0$ and $x \in D$, there exists $N_{\varepsilon, x}$ which may depend on x as well as ε such that when $q > p \geq N_{\varepsilon, x}$, $\left| \sum_{k=p}^q f_k(x) \right| < \varepsilon$. The series $\sum_{k=1}^{\infty} f_k$ converges uniformly on D if for every $\varepsilon > 0$ there exists N_{ε} such that if $q > p \geq N_{\varepsilon}$ then

$$\sup_{x \in D} \left| \sum_{k=p}^q f_k(x) \right| < \varepsilon \quad (6.10)$$

Proof: The first part follows from Theorem 6.1.8. The second part follows from observing the condition is equivalent to the sequence of partial sums forming a uniformly Cauchy sequence and then by Corollary 4.9.5, these partial sums converge uniformly to a function which is the definition of $\sum_{k=1}^{\infty} f_k$. ■

Is there an easy way to recognize when 6.10 happens? Yes, there is. It is called the Weierstrass M test.

Theorem 6.8.3 Let $\{f_n\}$ be a sequence of functions defined on D . Suppose there exists M_n such that $\sup\{|f_n(x)| : x \in D\} < M_n$ and $\sum_{n=1}^{\infty} M_n$ converges. Then $\sum_{n=1}^{\infty} f_n$ converges uniformly on D .

Proof: Let $z \in D$. Then letting $m < n$,

$$\left| \sum_{k=1}^n f_k(z) - \sum_{k=1}^m f_k(z) \right| \leq \sum_{k=m+1}^n |f_k(z)| \leq \sum_{k=m+1}^{\infty} M_k < \varepsilon$$

whenever m is large enough because of the assumption that $\sum_{n=1}^{\infty} M_n$ converges. Therefore, the sequence of partial sums is uniformly Cauchy on D and therefore, converges uniformly to $\sum_{k=1}^{\infty} f_k$ on D . ■

Theorem 6.8.4 *If $\{f_n\}$ is a sequence of functions defined on D which are continuous at z and $\sum_{k=1}^{\infty} f_k$ converges uniformly, then the function $\sum_{k=1}^{\infty} f_k$ must also be continuous at z .*

Proof: This follows from Theorem 4.9.3 applied to the sequence of partial sums of the above series which is assumed to converge uniformly to the function $\sum_{k=1}^{\infty} f_k$. ■

6.9 Exercises

1. Suppose $\{f_n\}$ is a sequence of decreasing positive functions defined on $[0, \infty)$ which converges pointwise to 0 for every $x \in [0, \infty)$. Can it be concluded that this sequence converges uniformly to 0 on $[0, \infty)$? Now replace $[0, \infty)$ with $(0, \infty)$. What can be said in this case assuming pointwise convergence still holds?
2. If $\{f_n\}$ and $\{g_n\}$ are sequences of functions defined on D which converge uniformly, show that if a, b are constants, then $af_n + bg_n$ also converges uniformly. If there exists a constant, M such that $|f_n(x)|, |g_n(x)| < M$ for all n and for all $x \in D$, show $\{f_n g_n\}$ converges uniformly. Let $f_n(x) \equiv 1/x$ for $x \in (0, 1)$ and let $g_n(x) \equiv (n-1)/n$. Show $\{f_n\}$ converges uniformly on $(0, 1)$ and $\{g_n\}$ converges uniformly but $\{f_n g_n\}$ fails to converge uniformly.
3. Show that if $x > 0$, $\sum_{k=0}^{\infty} \frac{x^k}{k!}$ converges uniformly on any interval of finite length.
4. Let $x \geq 0$ and consider the sequence $\left\{ \left(1 + \frac{x}{n}\right)^n \right\}$. Show this is an increasing sequence and is bounded above by $\sum_{k=0}^{\infty} \frac{x^k}{k!}$.
5. Show for every x, y real, $\sum_{k=0}^{\infty} \frac{(x+y)^k}{k!}$ converges and equals

$$\left(\sum_{k=0}^{\infty} \frac{y^k}{k!} \right) \left(\sum_{k=0}^{\infty} \frac{x^k}{k!} \right)$$

6. Consider the series $\sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!}$. Show this series converges uniformly on any interval of the form $[-M, M]$.
7. Formulate a theorem for a series of functions which will allow you to conclude the infinite series is uniformly continuous based on reasonable assumptions about the functions in the sum.

8. Find an example of a sequence of continuous functions 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 $(0, \infty)$.
9. Suppose $\{f_n\}$ is a sequence of real valued functions which converges uniformly to a continuous function f . Can it be concluded the functions f_n are continuous? Explain.
10. Let $h(x)$ be a bounded continuous function. Show the function $f(x) = \sum_{n=1}^{\infty} \frac{h(nx)}{n^2}$ is continuous.
11. Let S be a any countable subset of \mathbb{R} . This means S is actually the set of terms of a sequence. That is $S = \{s_n\}_{n=1}^{\infty}$. Show there exists a function f defined on \mathbb{R} 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 Theorem 6.8.4 and the Weierstrass M test.
12. By Theorem 4.10.3 there exists a sequence of polynomials converging uniformly to $f(x) = |x|$ on the interval $[-1, 1]$. Show there exists a sequence of polynomials, $\{p_n\}$ converging uniformly to f on $[-1, 1]$ which has the additional property that for all n , $p_n(0) = 0$.
13. If f is any continuous function defined on $[a, b]$, 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.
14. Sometimes a series may converge uniformly without the Weierstrass M test being applicable. Show $\sum_{n=1}^{\infty} (-1)^n \frac{x^2+n}{n^2}$ converges uniformly on $[0, 1]$ but does not converge absolutely for any $x \in \mathbb{R}$. To do this, it might help to use the partial summation formula, 6.7. Note that $\sum_{n=1}^{\infty} (-1)^n \frac{x^2+n}{n^2} = \sum_{n=1}^{\infty} (-1)^n \frac{x^2+n}{n} \left(\frac{1}{n}\right)$.

Chapter 7

The Integral

A more traditional treatment of the integral is described in the problems beginning with Problem 25 on Page 100. This approach is due to Darboux and is his description of the Riemann integral. The Riemann integral dates from the 1850's. It includes the case of continuous and piecewise continuous functions. However, the first integral to be adequate for considering continuous functions was due to Cauchy in 1820's who gave the first correct proof of the fundamental theorem of calculus which is normally credited to Newton and Leibniz. This is because the concept of what was meant by the integral was not precisely described until Cauchy. In fact, Cauchy's integral involved one sided sums and only worked well on continuous functions, but it was sufficient to give an acceptable proof for the fundamental theorem of calculus. For a complete discussion including Stieltjes integrals, a very important generalization, see my single variable advanced calculus book on my web site.

Although their understanding of the integral was incomplete, they found it as follows:

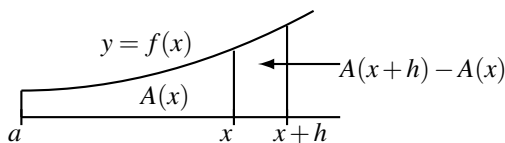
Procedure 7.0.1 To find $\int_a^b f(x) dx$ do the following:

1. Find $F(x)$ such that $F'(x) = f(x)$. Such an F is called an anti-derivative. Its derivative is an appropriate one sided derivative at the end points.
2. Then $\int_a^b f(x) dx \equiv F(b) - F(a)$

The above procedure to find " $\int_a^b f(x) dx$ ", was called the fundamental theorem of calculus. What they thought they were getting was a kind of infinite sum of the quantities $f(x) dx$, dx being an "infinitesimal" change in x , which is why it is denoted as $\int_a^b f(x) dx$, the long S symbolizing sum. Until Cauchy it was like this: We have something which we don't understand but it is like a sum and we can find it by using antiderivatives. It was like religious ritual. By contrast, Cauchy said exactly what he meant by the integral and showed that you could find it using antiderivatives. This is a big improvement.

However, defining the integral by the above Procedure, this is in fact a very interesting concept, because many applications can be formulated directly as a solution to an initial value problem from differential equations. This is a problem of the form $F'(x) = f(x)$, $F(0) = F_0$ where F is an unknown function and $F(0) = F_0$ is an initial condition. Here is a simple example, which is also the main historical motivation for the integral.

Consider $A(x)$ the area under the graph of a curve $y = f(x)$ as shown in the following picture between a and x .



Thus $A(x+h) - A(x) \in [f(x)h, f(x+h)h]$ and so

$$\frac{A(x+h) - A(x)}{h} \in [f(x), f(x+h)]$$

Then taking a limit as $h \rightarrow 0$, one obtains $A'(x) = f(x)$, $A(a) = 0$ and so one would have, from the above definition of the integral in terms of a procedure, $A(x) = \int_a^x f(t) dt$. This suggests that we should **define** the area under the graph of the curve between a and $x > a$ as this integral $\int_a^x f(t) dt$ which is computed as $A(x) - A(a)$ where $A' = f$. The argument is similar if f is decreasing.

Of course you should wonder whether you get the same thing for $\int_a^b f(x) dx$ if you use some other $\hat{F}(x)$ with $\hat{F}'(x) = f(x)$.

Proposition 7.0.2 *Procedure 7.0.1 is well defined in the sense that any F satisfying $F' = f$, on (a, b) with F continuous on $[a, b]$ yields the same answer for $\int_a^b f(x) dx$.*

Proof: Suppose both $F' = f$ and $\hat{F}' = f$. Then let $G(x) = F(x) - \hat{F}(x)$. For any $x, y \in [a, b]$, $G(x) - G(y) = G'(z)(x - y)$ for some z between x and y , this by the mean value theorem. However, $G'(z) = 0$ and so $G(x) = G(y)$ for every x, y . In particular, $F(b) - \hat{F}(b) - (F(a) - \hat{F}(a)) = 0$ which implies $F(b) - F(a) = \hat{F}(b) - \hat{F}(a)$. Also, $F(x) = \hat{F}(x) + C$ for some constant equal to the common value of G . ■

Example 7.0.3 *Find the area under the graph of the function $y = x^2$ where $0 \leq x \leq 2$.*

You find an antiderivative. One which works is $\frac{x^3}{3}$ because its derivative is x^2 . Then the desired area is $\frac{2^3}{3} - \frac{0^3}{3} = \frac{8}{3}$.

Example 7.0.4 *Find the area under $y = \sin x$ for $x \in [\frac{\pi}{2}, \pi]$.*

An antiderivative is $-\cos(x)$, so the desired area is $-\cos(\pi) - (-\cos(\frac{\pi}{2})) = 1$.

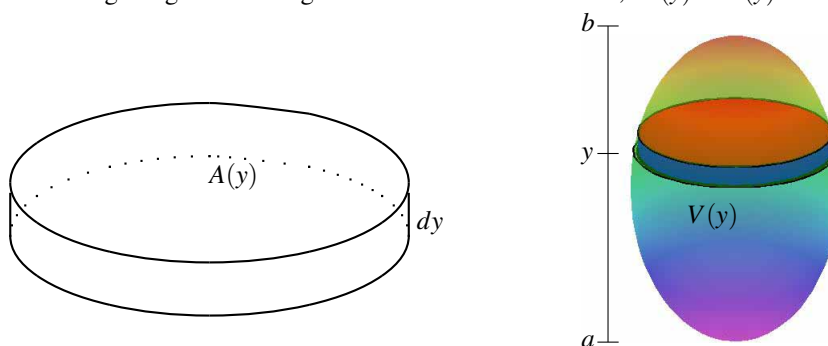
Of course the big question is whether there exists an antiderivative for an arbitrary continuous function. It turns out that the answer is yes, but sometimes we can't find it. However, there is one kind of function for which this is an easy problem. I know, for example that an antiderivative for $p(x) = 1 + x + 2x^2$ is $x + \frac{x^2}{2} + 2\frac{x^3}{3}$. Just look at it. It works. So is there an easy way to find an antiderivative? Yes for polynomials.

Lemma 7.0.5 *Let $p(x) = a_0 + a_1x + \cdots + a_{n-1}x^{n-1} + a_nx^n$. Then if $P(x) = a_0x + a_1\frac{x^2}{2} + \cdots + a_{n-1}\frac{x^n}{n} + a_n\frac{x^{n+1}}{n+1}$ it follows that $P'(x) = p(x)$. Thus if $p(x) = \sum_{k=0}^n a_kx^k$, then an antiderivative is $\sum_{k=0}^n a_k\frac{x^{k+1}}{k+1}$.*

Proof: This follows from the rules of differentiation. ■

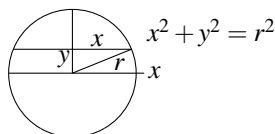
Definition 7.0.6 $\int f(x) dx$ denotes all functions F such that $F'(x) = f(x)$. Thus, from Proposition 7.0.2, $\int f(x) dx = F(x) + C$ where $F' = f$ and C is an arbitrary constant of integration.

There are many examples of how this can be used. Here is another one. Imagine a line next to a three dimensional solid as shown in the next picture. For each y between a and b , let $A(y)$ denote the area of the cross section of the solid obtained by intersecting this solid with a plane through y perpendicular to the indicated line. Then $\frac{V(y+h)-V(y)}{h} \approx \frac{hA(y)}{h} = A(y)$ the approximation getting better as h gets smaller. Thus in the limit, $V'(y) = A(y)$



and so the total volume of the solid between a and y , $V(y)$, satisfies the initial value problem $\frac{dV}{dy} = A(y)$, $V(a) = 0$. The volume of the solid is $V(b)$.

Example 7.0.7 Suppose a solid is obtained as a circular disk of radius r and center at $(0,0)$ is revolved about the y axis. Find the volume of the solid ball of radius r .



Here the line would be the y axis from $-r$ to r . For given y , you have $x^2 + y^2 = r^2$, so $x = \sqrt{r^2 - y^2}$ and this is the radius of a cross section located at y . Then by the above, $V'(y) = \pi(r^2 - y^2)$, $V(-r) = 0$. The volume is $\int_{-r}^r \pi(r^2 - y^2) dy$. An antiderivative is $\pi r^2 y - \pi \frac{y^3}{3}$ so the volume is $\left(\pi r^2 \cdot r - \pi \frac{r^3}{3}\right) - \left(\pi r^2 \cdot (-r) + \pi \frac{r^3}{3}\right) = \frac{4}{3} \pi r^3$. This is the volume of a ball of radius r .

The above procedure from the 1700's is how we find integrals. Thus we are finding solutions to an initial value problem and this will include many important applied problems in geometry and physics. However, there are significant theoretical questions, the most important being the existence of an antiderivative for a continuous function. These questions are resolved in the next section. After this, I will give a careful treatment of Darboux's formulation of the Riemannnn integral which will bring the understanding of the integral up to the 1850's. Expressing the integral as a limit of sums is the precise way of avoiding the fuzzy notion of infinitesimals which was the original idea of Leibniz. However, Leibniz's original idea is still very useful in formulating problems to be solved in terms of integrals.

At this point, the reader can do all of the do-able examples involving integrals except for finding antiderivatives, techniques for which are presented later. The following sections are theoretical in nature.

7.1 The Definition of the Integral from Antiderivatives

Next is the definition of what is meant by an oriented interval.

Definition 7.1.1 For the rest of this section, $[a, b]$ will denote the closed interval having end points a and b but a could be larger than b or smaller than b . It is written this way to indicate that there is a direction of motion from a to b which will be reflected by the definition of the integral given below. It is an “oriented interval”.

Definition 7.1.2 The integral of a continuous function defined on an oriented interval $[a, b]$ is defined by Procedure 7.0.1.

However, I need to verify that if f is continuous on $[a, b]$, then there is an antiderivative F in order to use that Procedure. This is the following lemma. It is a major result. Recall that for a function f defined on an interval $[a, b]$, $\|f\| \equiv \sup \{|f(x)| : x \in [a, b]\}$. Also recall that if f is a continuous function defined on $[a, b]$, then there exists a sequence of polynomials $\{p_n(x)\}$ for which $\|f - p_n\| \rightarrow 0$. This is by the Weierstrass approximation theorem of the chapter on continuous functions.

The message of the following lemma is that a continuous function on a closed interval has an antiderivative.

Lemma 7.1.3 Let f be a continuous, real valued function defined on $[a, b]$. Then there exists F such that $F'(x) = f(x)$ for all $x \in (a, b)$. At the end points, $F'(x)$ will refer to a one sided derivative.

Proof: Assume that $a < b$ in what follows. If not, simply switch a and b in the argument. Let $\{p_n\}$ be a sequence of polynomials for which $\|p_n - f\| \rightarrow 0$ and let $P'_n(x) = p_n(x)$ for all $x \in (a, b)$. By the mean value theorem,

$$\begin{aligned} & |P_n(x) - P_n(a) - (P_m(x) - P_m(a))| = \\ & |P_n(x) - P_m(x) - (P_n(a) - P_m(a))| \\ & = |(p_n(t) - p_m(t))(x - a)| \leq \|p_n - p_m\| |b - a| \\ & \leq (\|p_n - f\| + \|f - p_m\|) |b - a| \end{aligned}$$

The right side converges to 0 as $n, m \rightarrow \infty$ and so by completeness, there exists

$$F(x) = \lim_{n \rightarrow \infty} (P_n(x) - P_n(a)),$$

this for any choice of x . It remains to verify that $F'(x) = f(x)$. Say $x \in [a, b]$ and let $h > 0$. Then by the mean value theorem,

$$\frac{P_n(x+h) - P_n(x)}{h} = \frac{(P_n(x+h) - P_n(a)) - (P_n(x) - P_n(a))}{h} = p_n(t_{hn}) \quad (*)$$

for some $t_{hn} \in (x, x+h)$. By compactness, there is a subsequence, still denoted as t_{hn} for which $\lim_{n \rightarrow \infty} t_{hn} = t_h \in [x, x+h]$. Now

$$\begin{aligned} |p_n(t_{hn}) - f(t_h)| & \leq |p_n(t_{hn}) - f(t_{hn})| + |f(t_{hn}) - f(t_h)| \\ & \leq \|p_n - f\| + |f(t_{hn}) - f(t_h)| \end{aligned}$$

and so, letting $n \rightarrow \infty$, this shows, from continuity of f that $|p_n(t_{hn}) - f(t_h)| \rightarrow 0$. Taking a limit in $*$,

$$\frac{F(x+h) - F(x)}{h} = f(t_h), \quad t_h \in [x, x+h]$$

Now by continuity of f , we can take a limit of this as $h \rightarrow 0$ and obtain $F'(x) = f(x)$, where $F'(x)$ is a right derivative at $x = a$. For $x \in (a, b]$, the situation is exactly the same for when h is restrained to be negative. Indeed,

$$\frac{F(x+h) - F(x)}{h} = -\frac{F(x - (-h)) - F(x)}{-h} = \frac{F(x) - F(x-k)}{k}$$

where $k \equiv -h$ and so for $F'(x)$ the left derivative, it exists at each point of $(a, b]$ and equals $f(x)$. Also, the right derivative exists on $[a, b)$ and equals $f(x)$ and by similar reasoning, the left derivative exists on $(a, b]$ and equals $f(x)$. Thus F is continuous and $F'(x) = f(x)$ for $x \in (a, b)$. ■

Proposition 7.1.4 *The above integral is well defined for f continuous on $[a, b]$ 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 $\left| \int_a^b f dx \right| \leq \|f\| |b-a|$.
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 $F'(x) = f(x)$. Also,

$$\int_a^b (\alpha f(x) + \beta g(x)) dx = \alpha \int_a^b f(x) dx + \beta \int_a^b g(x) dx$$

$$\text{If } a < b, \text{ and } f(x) \geq 0, \text{ then } \int_a^b f dx \geq 0. \text{ Also } \left| \int_a^b f dx \right| \leq \left| \int_a^b |f| dx \right|.$$

4. $\int_a^b 1 dx = b - a$.

Proof: The integral is well defined by Lemma 7.1.3 and Proposition 7.0.2. Consider 1. Let $F'(x) = f(x)$, F as in Lemma 7.1.3 so

$$\int_a^b f(x) dx \equiv F(b) - F(a) = f(\hat{x})(b-a)$$

for some \hat{x} in the open interval determined by a, b . This is by the mean value theorem. Hence $\left| \int_a^b f dx \right| \leq \|f\| |b-a|$.

Now consider 2. Let $F' = f$ on a closed interval which contains all necessary intervals. Then from the definition,

$$\int_a^c f dx + \int_c^b f dx = F(c) - F(a) + F(b) - F(c) = F(b) - F(a) \equiv \int_a^b f(x) dx$$

Next consider 3. For $F(x) \equiv \int_a^x f(x) dx$, the definition says that $F(x) = G(x) - G(a)$ where $G'(x) = f(x)$ and so, since $G' = F'$, it follows that $F'(x) = f(x)$ with an appropriate

one sided derivative at the ends of the interval. Now let $F' = f, G' = g$. Then $\alpha f + \beta g = (\alpha F + \beta G)'$ and so

$$\begin{aligned} \int_a^b (\alpha f(x) + \beta g(x)) dx &\equiv (\alpha F + \beta G)(b) - (\alpha F + \beta G)(a) \\ &= \alpha F(b) + \beta G(b) - (\alpha F(a) + \beta G(a)) \\ &= \alpha(F(b) - F(a)) + \beta(G(b) - G(a)) \\ &\equiv \alpha \int_a^b f(x) dx + \beta \int_a^b g(x) dx \end{aligned}$$

If $f \geq 0, a < b$, then the mean value theorem implies that for $F' = f$, and some

$$t \in (a, b), F(b) - F(a) = \int_a^b f dx = f(t)(b - a) \geq 0.$$

Thus

$$\begin{aligned} \int_a^b (|f| - f) dx &\geq 0, \int_a^b (|f| + f) dx \geq 0 \text{ so} \\ \int_a^b |f| dx &\geq \int_a^b f dx \text{ and } \int_a^b |f| dx \geq - \int_a^b f dx \end{aligned}$$

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 $\left| \int_a^b f dx \right| \leq \left| \int_a^b |f| dx \right|$ even if $a > b$.

The last claim is obvious because an antiderivative of 1 is $F(x) = x$. ■

The change of variables theorem is available from the chain rule 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 also have the integration by parts formula from the product rule. Say $F' = f, G' = g$. Then from the product rule, $(FG)' = fG + gF$. In particular, if f, g are continuous on $[a, b]$,

$$F(b)G(b) - F(a)G(a) = \int_a^b f(t)G(t) dt + \int_a^b g(t)F(t) dt$$

These formulas are discussed more later.

Definition 7.1.5 A function $f : [a, b] \rightarrow \mathbb{R}$ is *piecewise continuous* if there is an ordered list of intermediate points z_i having an order consistent with $[a, b]$, meaning that $z_{i-1} - z_i$ has the same sign as $a - b$, $a = z_0, z_1, \dots, z_n = b$, called a *partition* of $[a, b]$, and functions f_i continuous on $[z_{i-1}, z_i]$ such that $f = f_i$ on (z_{i-1}, z_i) . For f piecewise continuous, define

$$\int_a^b f(t) dt \equiv \sum_{i=1}^n \int_{z_{i-1}}^{z_i} f_i(s) ds$$

If such a function f_i exists, then it is uniquely defined on $[z_{i-1}, z_i]$ as $f_i(z_i) \equiv \lim_{x \rightarrow z_i^-} f(x)$ with a similar definition for $f_i(z_{i-1})$.

Observation 7.1.6 *Note that this actually defines the integral even if the function has finitely many discontinuities and that changing the value of the function at finitely many points does not affect the integral.*

Of course this gives what appears to be a new definition because if f is continuous on $[a, b]$, then it is piecewise continuous for any such partition. However, it gives the same answer because, from this new definition, $\int_a^b f(t) dt = \sum_{i=1}^n (F(z_i) - F(z_{i-1})) = F(b) - F(a)$.

Suppose f, g are piecewise continuous. Then let $\{z_i\}_{i=1}^n$ include all the partition points of both of these functions. Then, since it was just shown that no harm is done by including more partition points, $\int_a^b \alpha f(t) + \beta g(t) dt \equiv$

$$\begin{aligned} \sum_{i=1}^n \int_{z_{i-1}}^{z_i} (\alpha f_i(s) + \beta g_i(s)) ds &= \sum_{i=1}^n \alpha \int_{z_{i-1}}^{z_i} f_i(s) ds + \sum_{i=1}^n \beta \int_{z_{i-1}}^{z_i} g_i(s) ds \\ &= \alpha \sum_{i=1}^n \int_{z_{i-1}}^{z_i} f_i(s) ds + \beta \sum_{i=1}^n \int_{z_{i-1}}^{z_i} g_i(s) ds = \alpha \int_a^b f(t) dt + \beta \int_a^b g(t) dt \end{aligned}$$

Also, the claim that $\int_a^b f dt = \int_a^c f dt + \int_c^b f dt$ is obtained exactly as before by considering all partition points on each integral preserving the order of the limits in the small intervals determined by the partition points.

Definition 7.1.7 *Let I be an interval. Then $\mathcal{X}_I(t)$ is 1 if $t \in I$ and 0 if $t \notin I$. Then a step function will be of the form $\sum_{k=1}^n c_k \mathcal{X}_{I_k}(t)$ where $I_k = [a_{k-1}, a_k]$ is an interval and $\{I_k\}_{k=1}^n$ are non-overlapping intervals whose union is an interval $[a, b]$ so $b - a = \sum_{k=1}^n (a_k - a_{k-1})$. Then, as explained above,*

$$\int_a^b \sum_{k=1}^n c_k \mathcal{X}_{I_k}(t) dt = \sum_{k=1}^n c_k \int_{a_{k-1}}^{a_k} 1 dt = \sum_{k=1}^n c_k (a_k - a_{k-1}).$$

Is this as general as a complete treatment of Riemann integration? No it is not. The 1800's version of the integral is presented in the next section which is due to Riemann and Darboux. However, this version is sufficiently general to include all cases which are typically of interest. It is also enough to build a theory of ordinary differential equations. Note that Proposition 7.1.4 says that $\int_a^b f(t) dt = G(b) - G(a)$ whenever $G' = f$. This proposition also proves the fundamental theorem of calculus discovered by Newton and Leibniz, $(\int_a^t f(s) ds)' = f(t)$ if f is continuous. Unlike what was done by Newton and Leibniz, this approach also includes a rigorous definition of what is meant by the integral. To summarize, here is the procedure for finding an integral of a piecewise continuous function.

Procedure 7.1.8 *Let f be continuous on $[a, b]$. To find $\int_a^b f(t) dt$, find an antiderivative of f, F . Then $\int_a^b f(t) dt = F(b) - F(a)$. If f is piecewise continuous and equals the continuous function f_k on (z_{k-1}, z_k) where f_k is continuous on $[z_{k-1}, z_k]$ and $a = z_0 < z_1 < z_2 < \dots < z_n = b$, then $\int_a^b f(x) dx \equiv \sum_{k=1}^n \int_{z_{k-1}}^{z_k} f_k(x) dx$.*

The main assertion of the above Proposition 7.1.4 is that for any f continuous, there exists a unique solution to the initial value problem $F'(t) = f(t)$, along with $F(a) = 0$ and it is $F(t) = \int_a^t f(x) dx$.

7.2 Uniform Convergence and the Integral

It turns out that uniform convergence is very agreeable in terms of the integral. The following is the main result.

Theorem 7.2.1 *Let f_n be continuous and converging uniformly to f on $[a, b]$, $a < b$. Then it follows f is also continuous and*

$$\int_a^b f dx = \lim_{n \rightarrow \infty} \int_a^b f_n dx$$

Proof: The uniform convergence implies f is also continuous. See Theorem 4.9.3. Therefore, $\int_a^b f dx$ exists. Using the triangle inequality and definition of $\|\cdot\|$ described earlier in conjunction with this theorem,

$$\begin{aligned} \left| \int_a^b f(x) dx - \int_a^b f_n(x) dx \right| &= \left| \int_a^b (f(x) - f_n(x)) dx \right| \\ &\leq \int_a^b |f(x) - f_n(x)| dx \leq \int_a^b \|f - f_n\| dx \leq \|f - f_n\| (b - a) \end{aligned}$$

which is given to converge to 0 as $n \rightarrow \infty$. ■

7.3 The Riemann Darboux Integral*

In the 1850's Riemann gave a completely satisfactory description of the integral. The one Cauchy gave had some problems. I will present Darboux's version of this integral and show that it is the same as the earlier one for continuous and piecewise continuous functions. I will also present the fundamental theorem of calculus from this integral. In this section, $[a, b]$ will represent the usual notion of an interval in which $a < b$.

Definition 7.3.1 *For f a bounded function, and $P = \{x_0, x_1, \dots, x_n\} \subseteq [a, b]$ where, $a = x_0 < \dots < x_n = b$, let*

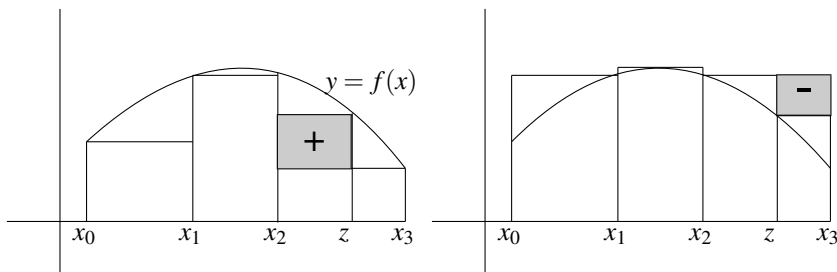
$$M_i(f) \equiv \sup \{f(x) : x \in [x_{i-1}, x_i]\}, \quad m_i(f) \equiv \inf \{f(x) : x \in [x_{i-1}, x_i]\}.$$

Then upper sums, $U(f, P)$ and lower sums $L(f, P)$ are defined as

$$U(f, P) \equiv \sum_{i=1}^n M_i(f) (x_i - x_{i-1}), \quad L(f, P) \equiv \sum_{i=1}^n m_i(f) (x_i - x_{i-1})$$

This collection of points is called a partition of $[a, b]$.

What happens when you add in more points in a partition? In this example a single additional point, labeled z has been added in.



Note how the lower sum got larger by the amount of the area in the shaded rectangle and the upper sum got smaller by the amount in the other shaded rectangle. In general this is the way it works and this is shown in the following lemma.

Lemma 7.3.2 *If $P \subseteq Q$ then*

$$U(f, Q) \leq U(f, P), \text{ and } L(f, P) \leq L(f, Q).$$

Proof: This is verified by adding in one point at a time. Thus let $P = \{x_0, \dots, x_n\}$ and let $Q = \{x_0, \dots, x_k, y, x_{k+1}, \dots, x_n\}$. Thus exactly one point y , is added between x_k and x_{k+1} . Now the term in the upper sum which corresponds to the interval $[x_k, x_{k+1}]$ in $U(f, P)$ is

$$\sup \{f(x) : x \in [x_k, x_{k+1}]\} (x_{k+1} - x_k) \quad (7.1)$$

and the terms which corresponds to the interval $[x_k, x_{k+1}]$ in $U(f, Q)$ are

$$\sup \{f(x) : x \in [x_k, y]\} (y - x_k) + \sup \{f(x) : x \in [y, x_{k+1}]\} (x_{k+1} - y) \quad (7.2)$$

$$\equiv M_1 (y - x_k) + M_2 (x_{k+1} - y) \quad (7.3)$$

All the other terms in the two sums coincide. Now

$$\sup \{f(x) : x \in [x_k, x_{k+1}]\} \geq \max(M_1, M_2)$$

and so the expression in 7.2 is no larger than

$$\begin{aligned} & \sup \{f(x) : x \in [x_k, x_{k+1}]\} (x_{k+1} - y) + \sup \{f(x) : x \in [x_k, x_{k+1}]\} (y - x_k) \\ &= \sup \{f(x) : x \in [x_k, x_{k+1}]\} (x_{k+1} - x_k), \end{aligned}$$

the term corresponding to the interval $[x_k, x_{k+1}]$ and $U(f, P)$. This proves the first part of the lemma pertaining to upper sums because if $Q \supseteq P$, one can obtain Q from P by adding in one point at a time and each time a point is added, the corresponding upper sum either gets smaller or stays the same and similarly, the resulting lower sum is no smaller. ■

Lemma 7.3.3 *If P and Q are two partitions, then*

$$L(f, P) \leq U(f, Q).$$

Proof: By Lemma 7.3.2,

$$L(f, P) \leq L(f, P \cup Q) \leq U(f, P \cup Q) \leq U(f, Q). \quad \blacksquare$$

Definition 7.3.4

$$\bar{I} \equiv \inf \{U(f, Q) \text{ where } Q \text{ is a partition}\}$$

$$\underline{I} \equiv \sup \{L(f, P) \text{ where } P \text{ is a partition}\}.$$

Note that \underline{I} and \bar{I} are well defined real numbers.

Theorem 7.3.5 $\underline{I} \leq \bar{I}$.

Proof: From Lemma 7.3.3,

$$\underline{I} = \sup\{L(f, P) \text{ where } P \text{ is a partition}\} \leq U(f, Q)$$

because $U(f, Q)$ is an upper bound to the set of all lower sums and so it is no smaller than the least upper bound. Therefore, since Q is arbitrary,

$$\begin{aligned} \underline{I} &= \sup\{L(f, P) \text{ where } P \text{ is a partition}\} \\ &\leq \inf\{U(f, Q) \text{ where } Q \text{ is a partition}\} \equiv \bar{I} \end{aligned}$$

where the inequality holds because it was just shown that \underline{I} is a lower bound to the set of all upper sums and so it is no larger than the greatest lower bound of this set. ■

Now here is the definition of the Darboux integral based on the observation that $\bar{I} \geq \underline{I}$.

Definition 7.3.6 For f a bounded function on $[a, b]$,

$$\begin{aligned} \bar{I} &\equiv \inf\{U(f, P) \text{ where } P \text{ is a partition}\}, \\ \underline{I} &\equiv \sup\{L(f, P) \text{ where } P \text{ is a partition}\}. \end{aligned}$$

Then f is integrable if $\bar{I} = \underline{I}$ and the Darboux integral is the common value of these. I will call \underline{I} the lower integral and \bar{I} the upper integral. Thus the function is integrable exactly when there is no gap between the upper and lower integrals.

Proposition 7.3.7 A bounded function f defined on $[a, b]$ is integrable if and only if for each $\varepsilon > 0$ there exists a partition P such that $U(f, P) - L(f, P) < \varepsilon$.

Proof: \Rightarrow In this case, the upper and lower integrals are equal and so $\underline{I} + \varepsilon/3 > \bar{I}$, $\bar{I} - \varepsilon/3 < \underline{I}$. Thus, there is a partition P such that $\underline{I} + \varepsilon/3 > U(f, P)$ and $\bar{I} - \varepsilon/3 < L(f, P)$. Therefore,

$$0 \leq U(f, P) - L(f, P) \leq \underline{I} + \varepsilon/3 - (\bar{I} - \varepsilon/3) < \varepsilon$$

\Leftarrow If there is some P such that $U(f, P) - L(f, P) < \varepsilon$, then $0 \leq \bar{I} - \underline{I} \leq U(f, P) - L(f, P) < \varepsilon$ and so, since ε is arbitrary, $\bar{I} - \underline{I} = 0$. There is no gap between the upper and lower integrals. ■

Proposition 7.3.8 If f is either increasing or decreasing on $[a, b]$, then f is integrable.

Proof: Suppose first that f is decreasing. There is no space between \bar{I} and \underline{I} because if $a = x_0 < x_1 < \dots < x_n = b$ where these points in the partition are equally spaced, then

$$\begin{aligned} \bar{I} - \underline{I} &\leq \sum_{k=1}^n f(x_{k-1})(x_k - x_{k-1}) - \sum_{k=1}^n f(x_k)(x_k - x_{k-1}) \\ &= \sum_{k=1}^n (f(x_{k-1}) - f(x_k)) \frac{b-a}{n} = (f(a) - f(b)) \frac{b-a}{n} \end{aligned}$$

Since n is arbitrary, it must be that $\bar{I} - \underline{I} = 0$. It is exactly similar for f increasing. You just take the upper sum by using the value of f at the right end of the interval and the lower sum by taking the value of f at the left end of the interval. ■

Corollary 7.3.9 Suppose $[a, b]$ is an interval and f is a bounded real valued function defined on this interval and that there is a partition $a = z_0 < z_1 < \cdots < z_n = b$ such that f is either increasing or decreasing on each sub interval $[z_{i-1}, z_i]$. Then $\int_a^b f dx$ exists. Thus all reasonable bounded functions are integrable.

Proof: Let \bar{I}_k and \bar{I}_k and I_k pertain to the interval $[z_{k-1}, z_k]$. Then these are equal and so there is a partition P of $[a, b]$ including all the z_k such that $U(f, P) - L(f, P) < \varepsilon$. You just consider an appropriate partition of $[z_{k-1}, z_k]$ making the difference between the upper and lower sums less than ε/n for each of these sub intervals. Thus there is no space between \bar{I} and \underline{I} because ε is arbitrary. ■

Theorem 7.3.10 Suppose a bounded real valued function f is integrable on $[a, c]$ and that $a < b < c$. Then the restrictions of this function to $[a, b]$ and $[b, c]$ are integrable on these intervals and in fact,

$$\int_a^b f dx + \int_b^c f dx = \int_a^c f dx$$

Proof: By assumption, there is a partition P_1 of $[a, b]$ and one for $[b, c]$ P_2 such that $U(f, P_1) - L(f, P_1) < \frac{\varepsilon}{2}$, $U(f, P_2) - L(f, P_2) < \frac{\varepsilon}{2}$. Thus if $P = P_1 \cup P_2$, then $U(f, P) - L(f, P) < \varepsilon$ and so there is no space between \bar{I} and \underline{I} . Thus the function is integrable on $[a, c]$ and also, using the partitions just described,

$$\begin{aligned} -\varepsilon &< L(f, P) - U(f, P) = L(f, P_1) + L(f, P_2) - U(f, P) \\ &\leq \int_a^b f dx + \int_b^c f dx - \int_a^c f dx \leq U(f, P_1) + U(f, P_2) - L(f, P) \\ &= U(f, P) - L(f, P) < \varepsilon \end{aligned}$$

Thus $\int_a^b f dx + \int_b^c f dx - \int_a^c f dx \in [-\varepsilon, \varepsilon]$ and ε is arbitrary so $\int_a^b f dx + \int_b^c f dx - \int_a^c f dx = 0$. ■

Definition 7.3.11 Define $\int_b^a f dx \equiv -\int_a^b f dx$.

Theorem 7.3.12 Let f be integrable on $[\min(p, q, r), \max(p, q, r)]$. Then $\int_p^q f dx + \int_q^r f dx = \int_p^r f dx$.

Proof: The case where $a < b < c$ was just done. Suppose $a < c < b$. Then from what was just done,

$$\int_a^c f dx + \int_c^b f dx = \int_a^b f dx$$

and so

$$\int_a^c f dx = \int_a^b f dx - \int_c^b f dx = \int_a^b f dx + \int_b^c f dx$$

Other cases are similar. ■

Theorem 7.3.13 If f is continuous on $[a, b]$, then f is integrable on $[a, b]$. Also, if f is integrable on $[a, b]$ and is changed at finitely many points, the resulting function is also integrable and has the same integral as f .

Proof: I will show there is no gap between the upper and lower integrals. Let $\varepsilon > 0$ be given. Let n be so large that if $|x - y| < 2\delta$, then $|f(x) - f(y)| < \varepsilon / ((b - a) + 1)$. Such a δ exists because f is uniformly continuous due to the fact that $[a, b]$ is compact. See Theorem 4.7.2. Now let n be so large that $\frac{b-a}{n} < \delta$. Then let $P = \{x_0, x_1, \dots, x_n\}$ be a uniform partition, each $x_k - x_{k-1} = \frac{b-a}{n}$. Then

$$U(f, P) - L(f, P) = \sum_{k=1}^n f(z_k)(x_k - x_{k-1}) - \sum_{k=1}^n f(w_k)(x_k - x_{k-1})$$

where $f(z_k)$ is the maximum value of f on $[x_{k-1}, x_k]$ and $f(w_k)$ the minimum value of f on $[x_{k-1}, x_k]$. Thus $|w_k - z_k| < 2\delta$ and so the above is no larger than

$$\sum_{k=1}^n \frac{\varepsilon}{(b-a)+1} \left(\frac{b-a}{n} \right) = (b-a) \frac{\varepsilon}{(b-a)+1} < \varepsilon$$

Thus, from Proposition 7.3.7, f is integrable.

Now consider the claim about an integrable function being changed at finitely many points. Let f be the integrable function. Let the finitely many points be z_1, z_2, \dots, z_r listed in order. Also let \hat{f} be the modified function and let

$$\begin{aligned} M &\equiv \sup(\max(|\hat{f}(x)|, |f(x)|), x \in [a, b]), \\ m &\equiv \inf(\min(-|\hat{f}(x)|, -|f(x)|), x \in [a, b]) \end{aligned}$$

Let $P \equiv \{x_0, x_1, \dots, x_n\}$ be a partition which contains all the finitely many points and suppose also that $|x_k - x_{k-1}| < \delta$ where $2(M - m)r\delta < \frac{\varepsilon}{2}$ and also

$$U(f, P) - L(f, P) < \frac{\varepsilon}{2}$$

Let the z_i be x_{k_i} . Then

$$\begin{aligned} U(\hat{f}, P) - L(\hat{f}, P) &\leq U(f, P) - L(f, P) \\ &\quad + (M - m) \sum_{i=1}^r (x_{k_i+1} - x_{k_i}) + (M - m) \sum_{i=1}^r (x_{k_i} - x_{k_i-1}) \\ &< \frac{\varepsilon}{2} + 2(M - m)r\delta < \varepsilon \end{aligned}$$

Since ε is arbitrary, this shows there is no gap between the upper and lower integrals and so \hat{f} is also integrable. ■

Corollary 7.3.14 *Every piecewise continuous function is integrable and if f is piecewise continuous and $f = f_k$ on (z_{k-1}, z_k) where f_k is continuous on $[z_{i-1}, z_i]$, then*

$$\int_a^b f(x) dx = \sum_{k=1}^r \int_{z_{k-1}}^{z_k} f_k(x) dx$$

Proof: From what was just shown, f is integrable on each $[z_{k-1}, z_k]$ because it equals a continuous function except maybe at the end points. Also, from induction and Theorem 7.3.12 f is integrable on $[a, b]$ and $\int_a^b f(x) dx = \sum_{k=1}^r \int_{z_{k-1}}^{z_k} f(x) dx = \sum_{k=1}^r \int_{z_{k-1}}^{z_k} f_k(x) dx$. ■

Now here is one version of the fundamental theorem of calculus.

Theorem 7.3.15 Let $f(x) = F'(x)$ for $x \in (a, b)$ and F is continuous on $[a, b]$. Also suppose f is integrable. Then $\int_a^b f(x) dx = F(b) - F(a)$.

Proof: There is a partition P such that $U(f, P) - L(f, P) < \varepsilon$. Then letting x_k denote the points of P in the usual way, by the mean value theorem, there exists $z_k \in (x_{k-1}, x_k)$ such that

$$\begin{aligned} F(b) - F(a) &= \sum_{k=1}^n F(x_k) - F(x_{k-1}) \\ &= \sum_{k=1}^n f(z_k)(x_k - x_{k-1}) \in [L(f, P), U(f, P)] \end{aligned}$$

an interval of length no more than ε . But also $\int_a^b f(x) dx$ is in this same interval and so

$$\left| (F(b) - F(a)) - \int_a^b f(x) dx \right| < \varepsilon$$

Since ε is arbitrary, this shows $\int_a^b f(x) dx = F(b) - F(a)$. ■

At this point this integral is seen to be a generalization of the earlier integral defined according to the fundamental theorem of calculus. Next consider functions of Riemann Darboux integrable functions.

Proposition 7.3.16 Suppose $H : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ satisfy

$$|H(x, y) - H(\hat{x}, \hat{y})| \leq K(|x - \hat{x}| + |y - \hat{y}|)$$

Then if $f, g \in R([a, b])$ it follows that $H(f, g) \in R([a, b])$.

Proof: By hypothesis and the Riemann criterion, there is a partition P such that for $h = f$ or g , $U(h, P) - L(h, P) < \frac{\varepsilon}{2K}$. Say $P = x_0 < x_1 < \dots < x_n$. Then consider

$$\sum_{i=1}^n (M_i(H(f, g)) - m_i(H(f, g)))(x_i - x_{i-1})$$

Say $H(f(z_i), g(z_i)) + \eta > M_i(H(f, g))$ and $H(f(w_i), g(w_i)) - \eta < m_i(H(f, g))$ where z_i, w_i are in $[x_{i-1}, x_i]$. Then

$$\begin{aligned} M_i(H(f, g)) - m_i(H(f, g)) &\leq H(f(z_i), g(z_i)) - (H(f(w_i), g(w_i))) + 2\eta \\ &\leq K(|f(z_i) - f(w_i)| + |g(z_i) - g(w_i)|) + 2\eta \\ &< K((M_i(f) - m_i(f)) + (M_i(g) - m_i(g))) + 2\eta \end{aligned}$$

Since η is arbitrary, it follows that

$$M_i(H(f, g)) - m_i(H(f, g)) \leq K((M_i(f) - m_i(f)) + (M_i(g) - m_i(g)))$$

and so

$$U(H(f, g), P) - L(H(f, g), P) < 2K \frac{\varepsilon}{2K} = \varepsilon$$

Since ε is arbitrary, this verifies that $H(f, g) \in R([a, b])$. ■

Note that $H(x, y) = \alpha x + \beta y$ satisfies the above conditions for α, β real numbers. Therefore, if $f, g \in R([a, b])$, it follows that the linear combination $\alpha f + \beta g$ is integrable. Similarly αf and βg are integrable. If f, g have values in some interval $[a, b]$ and $H : [a, b] \times [a, b] \rightarrow \mathbb{R}$ is only continuous, it could be shown that if f, g are integrable so is $H(f, g)$ but this is more trouble.

Lemma 7.3.17 *Let f, g be integrable. Then $\int_a^b (f + g)(x) dx = \int_a^b f(x) dx + \int_a^b g(x) dx$.*

Proof: For $x \in [x_{i-1}, x_i]$, $(f + g)(x) \geq m_i(f) + m_i(g)$ and so $m_i(f + g) \geq m_i(f) + m_i(g)$. Similarly $M_i(f + g) \leq M_i(f) + M_i(g)$. Therefore,

$$U(f + g, P) \leq U(f, P) + U(g, P), \quad L(f + g, P) \geq L(f, P) + L(g, P)$$

Let $U(f, P) - L(f, P) < \varepsilon$ and $U(g, P) - L(g, P) < \varepsilon$. Then

$$\begin{aligned} \int_a^b (f + g)(x) dx &\in [L(f + g, P), U(f + g, P)] \\ &\subseteq [L(f, P) + L(g, P), U(f, P) + U(g, P)] \\ \int_a^b f(x) dx + \int_a^b g(x) dx &\in [L(f, P) + L(g, P), U(f, P) + U(g, P)] \end{aligned}$$

Thus both $\int_a^b (f + g)(x) dx$ and $\int_a^b f(x) dx + \int_a^b g(x) dx$ are in an interval of length 2ε and so $\left| \int_a^b (f + g)(x) dx - \left(\int_a^b f(x) dx + \int_a^b g(x) dx \right) \right| < 2\varepsilon$. Since ε is arbitrary, this proves the lemma. ■

Lemma 7.3.18 *Let $\alpha \geq 0$. Then $\int_a^b \alpha f(x) dx = \alpha \int_a^b f(x) dx$.*

Proof: It is routine to verify that $U(\alpha f, P) = \alpha U(f, P)$, $L(\alpha f, P) = \alpha L(f, P)$. Let $U(f, P) - L(f, P) < \varepsilon$. Therefore,

$$\begin{aligned} \int_a^b \alpha f(x) dx &\in [L(\alpha f, P), U(\alpha f, P)] = [\alpha L(f, P), \alpha U(f, P)] \\ \alpha \int_a^b f(x) dx &\in [\alpha L(f, P), \alpha U(f, P)] \end{aligned}$$

Thus both $\int_a^b \alpha f(x) dx$ and $\alpha \int_a^b f(x) dx$ are in an interval of length $\alpha\varepsilon$ so since ε is arbitrary, this proves the lemma. ■

Proposition 7.3.19 *Let f, g be integrable. Then if α, β are real numbers, then the following holds for the integrals: $\int_a^b (\alpha f + \beta g)(x) dx = \alpha \int_a^b f(x) dx + \beta \int_a^b g(x) dx$.*

Proof: I want to show that $\int_a^b \alpha f(x) dx = \alpha \int_a^b f(x) dx$ regardless of whether α is nonnegative. By the first lemma,

$$\int_a^b (|\alpha| - \alpha) f(x) dx + \int_a^b \alpha f(x) dx = \int_a^b |\alpha| f(x) dx$$

and now, by the second lemma and subtracting $|\alpha| \int_a^b f(x) dx$,

$$\begin{aligned} (|\alpha| - \alpha) \int_a^b f(x) dx + \int_a^b \alpha f(x) dx &= \int_a^b |\alpha| f(x) dx \\ -\alpha \int_a^b f(x) dx + \int_a^b \alpha f(x) dx &= 0 \end{aligned}$$

so this shows one can factor out any real number.

Now from what was just shown and the lemmas,

$$\begin{aligned}\int_a^b (\alpha f + \beta g)(x) dx &= \int_a^b \alpha f(x) dx + \int_a^b \beta f(x) dx \\ &= \alpha \int_a^b f(x) dx + \beta \int_a^b f(x) dx \quad \blacksquare\end{aligned}$$

Proposition 7.3.20 *If $a < b$, Then $\int_a^b |f(x)| dx \geq \left| \int_a^b f(x) dx \right|$. Here f is assumed integrable.*

Proof: From the definition, if $a < b$, then $\int_a^b f(x) dx \geq 0$ if $f(x) \geq 0$ for all x . Now from Proposition 7.3.16, if f is integrable, so is $|f|$. Then

$$\begin{aligned}\int_a^b (|f(x)| - f(x)) dx &\geq 0 \text{ so } \int_a^b |f(x)| dx \geq \int_a^b f(x) dx \\ \int_a^b (|f(x)| + f(x)) dx &\geq 0 \text{ so } \int_a^b |f(x)| dx \geq -\int_a^b f(x) dx\end{aligned}$$

which implies that for $a < b$, $\int_a^b |f(x)| dx \geq \left| \int_a^b f(x) dx \right|$. ■

Definition 7.3.21 *Suppose f is a function and P is a partition $P = x_0 = a < x_1 < \dots < x_n = b$. A Riemann sum is of the form $\sum_{k=1}^n f(z_k)(x_k - x_{k-1})$ where $z_k \in [x_{k-1}, x_k]$. Thus every such Riemann sum is between $U(f, P)$ and $L(f, P)$ and so, if f is integrable, every such Riemann sum can be considered an approximation to $\int_a^b f(x) dx$.*

7.4 Exercises

1. Let $f(x) = \sin(1/x)$ for $x \in (0, 1]$ and let $f(0) = 0$. Show that f is Riemann Darboux integrable. Is f piecewise continuous?
2. Show that if f is Riemann Darboux integrable and if $F(x) = \int_a^x f(t) dt$, then $F'(x) = f(x)$ at every point x where f is continuous.
3. Show that if f_n is Riemann Darboux integrable on $[a, b]$, and $f_n \rightarrow f$ uniformly on $[a, b]$, then f is also Riemann Darboux integrable and $\int_a^b f_n(x) dx \rightarrow \int_a^b f(x) dx$.
4. The first order linear initial value problem for the unknown function y is of the form $y'(x) + p(x)y(x) = q(x)$, $y(0) = y_0$ where y_0 is given and $p(x), q(x)$ are given continuous functions. How do you find y in this problem? This will be discussed in this problem. Incidentally, this is the most important equation in differential equations and properly understood includes almost the entire typical undergraduate differential equations course. Let $P(x) \equiv \int_0^x p(t) dt$. Then multiply both sides by $\exp(P(x))$ and when you do, show that you obtain $\frac{d}{dx}(\exp(P(x))y(x)) = q(x)\exp(P(x))$. Now explain why, when you take an integral of both sides, you get $\exp(P(x))y(x) - y_0 = \int_0^x q(t)\exp(P(t)) dt$. Then

$$y(x) = \exp(-P(x))y_0 + \exp(-P(x)) \int_0^x q(t)\exp(P(t)) dt$$

5. One of the most important inequalities in differential equations is Gronwall's inequality. You have $u(t) \leq u_0 + K \int_0^t u(s) ds$, $t \geq 0$ where $t \rightarrow u(t)$ is some continuous function usually nonnegative. Then you can conclude that $u(t) \leq u_0 e^{Kt}$. Explain why this is so. **Hint:** Let $w(t) = \int_0^t u(s) ds$ and write the inequality in terms of w and its derivatives. Then use the technique of the previous problem involving integrating factors.
6. A function f satisfies a Lipschitz condition if $|f(x) - f(y)| \leq K|x - y|$. A standard initial value problem is to find a function of t denoted as y such that $y'(t) = f(y(t))$, $y(0) = y_0$ where y_0 is a given value called an initial condition. Show that this initial value problem has a solution if and only if there is a solution to the integral equation

$$y(t) = y_0 + \int_0^t f(y(s)) ds, \quad t \geq 0 \quad (7.4)$$

Hint: This is an application of theorems about continuity and the fundamental theorem of calculus.

7. Letting f be Lipschitz continuous as in 7.4, use Gronwall's inequality of Problem 5, to show there is at most one function y which is a solution to the integral equation 7.4. **Hint:** If y, \hat{y} both work, explain why $|y(t) - \hat{y}(t)| \leq \int_0^t K |y(s) - \hat{y}(s)| ds$. Also give a continuous dependence theorem in the case that you have y, \hat{y} solutions to

$$y(t) = y_0 + \int_0^t f(y(s)) ds, \quad t \geq 0 \quad \text{and} \quad \hat{y}(t) = \hat{y}_0 + \int_0^t f(\hat{y}(s)) ds, \quad t \geq 0$$

respectively. Verify $|y_0 - \hat{y}_0| e^{Kt} \geq |y(t) - \hat{y}(t)|$.

8. In fact, show there exists a solution to the initial value problem which is to find y such that $y(t) = y_0 + \int_0^t f(s, y(s)) ds$ under these conditions for $t \in [0, T]$. **Hint:** Use Picard iteration. Let $y_0(t) = y_0$ and if $y_n(t)$ has been obtained, let $y_{n+1}(t) = y_0 + \int_0^t f(s, y_n(s)) ds$ and show, using the Weierstrass M test on a telescoping series that this sequence converges uniformly to a continuous function y which is the solution to the integral equation and hence the initial value problem.
9. Give an example of piecewise continuous nonnegative functions f_n defined on $[0, 1]$ which converge pointwise to 0 but $\int_0^1 f_n(x) dx = 1$ for all n . This will show how uniform convergence or something else in addition to pointwise convergence is needed to get a conclusion like that in Theorem 7.2.1.
10. Let $F(x) = \int_{x^2}^{x^3} \frac{t^5 + 7}{t^7 + 87t^6 + 1} dt$. Find $F'(x)$.
11. Let $F(x) = \int_2^x \frac{1}{1+t^4} dt$. Sketch a graph of F and explain why it looks the way it does.
12. Let $a, b > 0$ and $F(x) = \int_0^{ax} \frac{1}{a^2 + t^2} dt + \int_b^{a/x} \frac{1}{a^2 + t^2} dt$. Show that F is a constant. **Hint:** Use the fundamental theorem of calculus.
13. Here is a function:

$$f(x) = \begin{cases} x^2 \sin\left(\frac{1}{x^2}\right) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

Show this function has a derivative at every point of \mathbb{R} . Does it make any sense to write $\int_0^1 f'(x) dx = f(1) - f(0) = f(1)$? Explain. Does this somehow contradict the fundamental theorem of calculus?

14. $\sum_{k=1}^n f(x_{k-1})(x_k - x_{k-1})$, $\sum_{k=1}^n f(x_k)(x_k - x_{k-1})$ are called left and right sums. Also suppose that all partitions have the property that $x_k - x_{k-1}$ is a constant, $(b-a)/n$ so the points in the partition are equally spaced, and define the integral to be the number these right and left sums get close to as n gets larger and larger. Show that for f given as 1 on rational numbers and 0 on irrational numbers, $\int_0^x f(t) dt = x$ if x is rational and $\int_0^x f(t) dt = 0$ if x is irrational. It turns out that the correct answer should always equal zero for that function, regardless of whether x is rational. This illustrates why this method of defining the integral in terms of left and right sums is terribly flawed. Show that even though this is the case, it makes no difference if f is continuous. This integral was used by Cauchy in the early 1800's. He considered one sided sums for continuous functions and ended up giving the first complete proof of the fundamental theorem of calculus.
15. Suppose f is a bounded function on $[0, 1]$ and for each $\varepsilon > 0$, $\int_\varepsilon^1 f(x) dx$ exists. Can you conclude $\int_0^1 f(x) dx$ exists? You need to be in the situation of the 1800's integral to do this problem.
16. Suppose f is a continuous function on $[a, b]$ and $\int_a^b f^2(x) dx = 0$. Show that then $f(x) = 0$ for all x .
17. Let f be Riemann integrable on $[0, 1]$. Show that $x \rightarrow \int_0^x f(t) dt$ is continuous. **Hint:** It is always assumed that Riemann integrable functions are bounded.
18. Define $F(x) \equiv \int_0^x \frac{1}{1+t^2} dt$. Of course $F(x) = \arctan(x)$ as mentioned above but just consider this function in terms of the integral. Sketch the graph of F using only its definition as an integral. Show there exists a constant M such that $-M \leq F(x) \leq M$. Next explain why $\lim_{x \rightarrow \infty} F(x)$ exists and show this limit equals $-\lim_{x \rightarrow -\infty} F(x)$.
19. In Problem 18 let the limit defined there be denoted by $\pi/2$ and define $T(x) \equiv F^{-1}(x)$ for $x \in (-\pi/2, \pi/2)$. Show $T'(x) = 1 + T(x)^2$ and $T(0) = 0$. As part of this, you must explain why $T'(x)$ exists. For $x \in [0, \pi/2]$ let $C(x) \equiv 1/\sqrt{1 + T(x)^2}$ with $C(\pi/2) = 0$ and on $[0, \pi/2]$, define $S(x)$ by $\sqrt{1 - C(x)^2}$. Show both $S(x)$ and $C(x)$ are differentiable on $[0, \pi/2]$ and satisfy $S'(x) = C(x)$ and $C'(x) = -S(x)$. Find the appropriate way to define $S(x)$ and $C(x)$ on all of $[0, 2\pi]$ in order that these functions will be $\sin(x)$ and $\cos(x)$ and then extend to make the result periodic of period 2π on all of \mathbb{R} . Note this is a way to define the trig. functions which is independent of plane geometry and also does not use power series. See the book by Hardy (If I remember correctly), [19] for this approach.
20. Let $p, q > 1$ and satisfy $\frac{1}{p} + \frac{1}{q} = 1$. Let $x = t^{p-1}$. Then solving for t , you get $t = x^{1/(p-1)} = x^{q-1}$. Explain this. Now let $a, b \geq 0$. Sketch a picture to show why

$$\int_0^b x^{q-1} dx + \int_0^a t^{p-1} dt \geq ab.$$

Now do the integrals to obtain a very important inequality $\frac{b^q}{q} + \frac{a^p}{p} \geq ab$. When will equality hold in this inequality?

21. Suppose f, g are two Riemann integrable functions on $[a, b]$. Verify Holder's inequality.

$$\int_a^b |f| |g| dx \leq \left(\int_a^b |f|^p dx \right)^{1/p} \left(\int_a^b |g|^q dx \right)^{1/q}$$

Hint: Do the following. Let $A = \left(\int_a^b |f|^p dx \right)^{1/p}$, $B = \left(\int_a^b |g|^q dx \right)^{1/q}$. Then let $a = \frac{|f|}{A}$, $b = \frac{|g|}{B}$ and use the wonderful inequality of Problem 20.

22. If F, G are antiderivatives for f, g on \mathbb{R} , show $F(x) = G(x) + C$ for some constant, C . Use this to give a proof of the fundamental theorem of calculus which has for its conclusion $\int_a^b f(t) dt = G(b) - G(a)$ where $G'(x) = f(x)$. Use the version of the fundamental theorem of calculus which says that $(\int_a^x f(t) dt)' = f(x)$ for f continuous.
23. Suppose f and g are continuous functions on $[a, b]$ and that $g(x) \neq 0$ on (a, b) . Show there exists $c \in [a, b]$ such that $f(c) \int_a^b g(x) dx = \int_a^b f(x) g(x) dx$. **Hint:** Define $m \equiv \min \{f(x) : x \in [a, b]\}$, $M \equiv \max \{f(x) : x \in [a, b]\}$. Now consider

$$\frac{\int_a^b f(x) g(x) dx}{\int_a^b g(x) dx} \quad \text{or} \quad \frac{\int_a^b f(x) (-g(x)) dx}{\int_a^b (-g(x)) dx}$$

Argue that one of these quotients is between m and M . Use intermediate value theorem.

24. A differentiable function f defined on $(0, \infty)$ satisfies the following conditions.

$$f(xy) = f(x) + f(y), \quad f'(1) = 1.$$

Find f and sketch its graph.

25. There is a general procedure for constructing methods of approximate integration. Consider $[0, 1]$ and divide this interval into n equal pieces determined by $\{x_0, \dots, x_n\}$ where $x_i - x_{i-1} = 1/n$ for each i . The approximate integration scheme for a function f , will be of the form

$$\left(\frac{1}{n} \right) \sum_{i=0}^n c_i f_i \approx \int_0^1 f(x) dx$$

where $f_i = f(x_i)$ and the constants, c_i are chosen in such a way that the above sum gives the exact answer for $\int_0^1 f(x) dx$ where $f(x) = 1, x, x^2, \dots, x^n$. When this has been done, change variables to write

$$\begin{aligned} \int_a^b f(y) dy &= (b-a) \int_0^1 f(a + (b-a)x) dx \\ &\approx \frac{b-a}{n} \sum_{i=1}^n c_i f \left(a + (b-a) \left(\frac{i}{n} \right) \right) = \frac{b-a}{n} \sum_{i=1}^n c_i f_i \end{aligned}$$

where $f_i = f \left(a + (b-a) \left(\frac{i}{n} \right) \right)$. Show that when $n = 1$, you get an approximation with trapezoids and when $n = 2$ you get an approximation with second degree polynomials. This is called Simpson's rule. Show also that if this integration scheme is

applied to any polynomial of degree 3 the result will be exact. That is,

$$\frac{1}{2} \left(\frac{1}{3} f_0 + \frac{4}{3} f_1 + \frac{1}{3} f_2 \right) = \int_0^1 f(x) dx$$

whenever $f(x)$ is a polynomial of degree three. Show that if f_i are the values of f at a , $\frac{a+b}{2}$, and b with $f_1 = f\left(\frac{a+b}{2}\right)$, it follows that the above formula gives $\int_a^b f(x) dx$ exactly whenever f is a polynomial of degree three.

26. Let f have four continuous derivatives on $[x_{i-1}, x_{i+1}]$ where $x_{i+1} = x_{i-1} + 2h$ and $x_i = x_{i-1} + h$. Show using Problem 17, there exists a polynomial of degree three, $p_3(x)$, such that

$$f(x) = p_3(x) + \frac{1}{4!} f^{(4)}(\xi)(x - x_i)^4$$

Now use Problem 25 to conclude

$$\left| \int_{x_{i-1}}^{x_{i+1}} f(x) dx - \left(\frac{hf_{i-1}}{3} + \frac{hf_i}{3} + \frac{hf_{i+1}}{3} \right) \right| < \frac{M}{4!} \frac{2h^5}{5},$$

where M satisfies, $M \geq \max \left\{ \left| f^{(4)}(t) \right| : t \in [x_{i-1}, x_i] \right\}$. You will approximate the integral with

$$\sum_{i=0}^{m-1} \left(\frac{hf(x_{2ih})}{3} + \frac{4hf(x_{2ih+h})}{3} + \frac{hf(x_{(2i+2)h})}{3} \right)$$

Note how this does an approximation for $0, h, 2h$, then from $2h, 3h, 4h$, etc. Denote this as $S(a, b, f, 2m)$ denote the approximation to $\int_a^b f(x) dx$ obtained from Simpson's rule using $2m+1$ equally spaced points with x_0 at the left. Show

$$\left| \int_a^b f(x) dx - S(a, b, f, 2m) \right| < \frac{M}{1920} (b-a)^5 \frac{1}{m^4}$$

where $M \geq \max \left\{ \left| f^{(4)}(t) \right| : t \in [a, b] \right\}$. Better estimates are available in numerical analysis books but these also have the error in the form $C(1/m^4)$.

27. Suppose f_n converges uniformly to f on $[a, b]$ and that $f_n \in R([a, b])$. Show $f \in R([a, b])$ and that $\int_a^b f(x) dx = \lim_{n \rightarrow \infty} \int_a^b f_n(x) dx$. That is, the uniform limit of Riemann integrable functions is Riemann integrable.
28. Find a power series to approximate $\ln(1-x)$ about 0 and show the remainder term converges to 0 if $|x| < 1$.
29. Find a power series to approximate $\ln(1+x)$ about 0 and show the remainder term converges to 0 if $|x| < 1$.
30. Give a series which will approximate $\ln\left(\frac{1+x}{1-x}\right)$ whenever $|x| < 1$. Show that for any $r > 0$, there is $x, |x| < 1$ such that $\frac{1+x}{1-x} = r$. Explain why the partial sums of the series will converge to $\ln r$.

31. A two dimensional shape S in a plane has area A and a cone is formed from drawing all lines from a point in S to a single point at height h above S . By similar triangles, the linear dimensions of similar shapes at height h in a plane parallel to the given plane are $\frac{h-y}{h}$ times the corresponding ones in the plane at the base. Thus if $A(y)$ is the area of the cross section in the plane at height y corresponding to S , it follows that the cross section at height y has area $A(y) = \frac{A}{h^2} (h-y)^2$. Find the volume of the cone. **Hint:** Show $\int (h-y)^2 dy = -\frac{(h-y)^3}{3} + C$ and use this. This includes pyramids, tetrahedra, circular cones, etc.

7.5 Videos

[Riemann integral](#)

Chapter 8

Methods for Finding Antiderivatives

There are methods for finding antiderivatives. These standard methods are recipes. They don't always work but they are the best we have. It turns out you can't find antiderivatives in terms of elementary functions in a routine way as you can with finding derivatives. It is a skill, not a substantial part of mathematics.

8.1 The Method of Substitution

I will illustrate the method of substitution by the use of examples. The method is somewhat formal. However, it works and you can check the answers obtained. Ultimately it is based on the chain rule for derivatives.

Example 8.1.1 Find $\int \sqrt[3]{2x+7}x dx$.

In this example $u = 2x + 7$ so that $du = 2dx$. Then

$$\begin{aligned}\int \sqrt[3]{2x+7}x dx &= \int \sqrt[3]{\overbrace{u-7}^x \overbrace{\frac{1}{2}}^{dx}} du = \int \left(\frac{1}{4}u^{4/3} - \frac{7}{4}u^{1/3} \right) du \\ &= \frac{3}{28}u^{7/3} - \frac{21}{16}u^{4/3} + C = \frac{3}{28}(2x+7)^{7/3} - \frac{21}{16}(2x+7)^{4/3} + C\end{aligned}$$

Example 8.1.2 Find $\int xe^{x^2} dx$.

Define a new variable $u = x^2$. Then $\frac{du}{dx} = 2x$ and so $du = 2x dx$ and $x dx = \frac{1}{2} du$. Then in terms of u the above integral is $\frac{1}{2} \int e^u du = \frac{1}{2} e^u + C$. Now substitute in what u equals in terms of x . This yields $\frac{1}{2} e^{x^2} + C$. Next check your work. Take the derivative of what you think the answer is and verify that it really is an antiderivative.

Example 8.1.3 $\int \sin(x) \cos(x) \sqrt{1 + \sin^2(x)} dx$

This can be done as follows. Let $u = 1 + \sin^2(x)$ so $du = 2 \sin(x) \cos(x) dx$ and the integral in the example simplifies to

$$\frac{1}{2} \int \sqrt{u} du = \frac{1}{3} u^{\frac{3}{2}} + C = \frac{1}{3} (1 + \sin^2(x))^{\frac{3}{2}} + C$$

You might try letting $u = \sin^2(x)$. It will also work but will likely take longer.

This illustrates that you are not always sure what substitution to use, but ultimately this method depends on the chain rule.

$$\int f(g(x)) g'(x) dx = F(g(x)) + C, \quad (8.1)$$

where $F'(y) = f(y)$. Here is another example.

Example 8.1.4 Find $\int x 3^{x^2} dx$

Let $u = 3^{x^2}$ so that $\frac{du}{dx} = 2x \ln(3) 3^{x^2}$ and $\frac{du}{2 \ln(3)} = x 3^{x^2} dx$. Thus

$$\int x 3^{x^2} dx = \frac{1}{2 \ln(3)} \int du = \frac{1}{2 \ln(3)} [u + C] = \frac{1}{2 \ln(3)} 3^{x^2} + \left(\frac{1}{2 \ln(3)} \right) C$$

Since the constant is an arbitrary constant, this is written as $\frac{1}{2 \ln(3)} 3^{x^2} + C$.

Example 8.1.5 Find $\int \cos^2(x) dx$

Recall that $\cos(2x) = \cos^2(x) - \sin^2(x)$ and $1 = \cos^2(x) + \sin^2(x)$. Then subtracting and solving for $\cos^2(x)$,

$$\cos^2(x) = \frac{1 + \cos(2x)}{2}.$$

Therefore,

$$\int \cos^2(x) dx = \int \frac{1 + \cos(2x)}{2} dx$$

Now letting $u = 2x$, $du = 2dx$ and so

$$\int \cos^2(x) dx = \int \frac{1 + \cos(u)}{4} du = \frac{1}{4} u + \frac{1}{4} \sin u + C = \frac{1}{4} (2x + \sin(2x)) + C.$$

Also $\int \sin^2(x) dx = -\frac{1}{2} \cos x \sin x + \frac{1}{2} x + C$ which is left as an exercise. This trick involving a trig. identity is almost the only way to do these.

Example 8.1.6 Find $\int \tan(x) dx$

Let $u = \cos x$ so that $du = -\sin(x) dx$. Then writing the antiderivative in terms of u , this becomes $\int \frac{-1}{u} du$. At this point, recall that $(\ln|u|)' = 1/u$. Thus this antiderivative is $-\ln|u| + C = \ln|u^{-1}| + C$ and so $\int \tan(x) dx = \ln|\sec x| + C$.

This illustrates a general procedure.

Procedure 8.1.7 $\int \frac{f'(x)}{f(x)} dx = \ln|f(x)| + C.$

This follows from the chain rule and the derivative of $x \rightarrow \ln|x|$.

Example 8.1.8 Find $\int \sec(x) dx$.

This is usually done by a trick. You write as $\int \frac{\sec(x)(\sec(x)+\tan(x))}{(\sec(x)+\tan(x))} dx$ and note that the numerator of the integrand is the derivative of the denominator. Thus

$$\int \sec(x) dx = \ln |\sec(x) + \tan(x)| + C.$$

Example 8.1.9 Find $\int \csc(x) dx$.

This is done like the antiderivatives for the secant. $\frac{d}{dx} \csc(x) = -\csc(x) \cot(x)$ and $\frac{d}{dx} \cot(x) = -\csc^2(x)$. Write the integral as

$$-\int \frac{-\csc(x)(\cot(x) + \csc(x))}{(\cot(x) + \csc(x))} dx = -\ln |\cot(x) + \csc(x)| + C.$$

Definition 8.1.10 Let $r(t)$ give a point on \mathbb{R} and regard t as time. This is called the position of the point. Then the velocity of the point is defined as $v(t) = r'(t)$. The acceleration is defined as the derivative of the velocity. Thus the acceleration is $a(t) \equiv r''(t)$.

Example 8.1.11 Let the velocity $v(t)$ of a point be given by $t^2 + 1$ and suppose the point is at 1 when $t = 0$. Find the position of the point.

Let the position of the point be $r(t)$. Then by definition of velocity, $r'(t) = t^2 + 1$ so $r(t) = \frac{t^3}{3} + t + C$. Now C must be determined. It is assumed that $r(0) = 1$. Therefore, $C = 1$ and so $r(t) = \frac{t^3}{3} + t + 1$.

Example 8.1.12 The acceleration of an object is given by $a(t) = t + 1$. When $t = 0$, the velocity is 1 and the position is 2. Determine the position.

It is given that $r''(t) = t + 1, r'(0) = 1$. Therefore, $r'(t) = \frac{t^2}{2} + t + 1$. Then $r(t) = \frac{t^3}{6} + \frac{t^2}{2} + t + 2$.

8.2 Exercises

1. Find the indicated antiderivatives.

(a) $\int \frac{x}{\sqrt{2x-3}} dx$

(b) $\int x(3x^2 + 6)^5 dx$

(c) $\int x \sin(x^2) dx$

(d) $\int \sin^3(2x) \cos(2x) dx$

(e) $\int \frac{1}{\sqrt{1+4x^2}} dx$ **Hint:** Remember the \sinh^{-1} function and its derivative.

2. Solve the initial value problems. There is an unknown function y and you are given its derivative and its value at some point.

- (a) $\frac{dy}{dx} = \frac{x}{\sqrt{2x-3}}, y(2) = 1$ (d) $y'(x) = \frac{1}{\sqrt{1+3x^2}}, y(1) = 1$
- (b) $\frac{dy}{dx} = 5x(3x^2+6)^5, y(0) = 3$ (e) $y'(x) = \sec(x), y(0) = 3$
- (c) $\frac{dy}{dx} = 3x^2 \sin(2x^3), y(1) = 1$ (f) $y'(x) = x \csc(x^2), y(1) = 1$
3. An object moves on the x axis having velocity equal to $\frac{3t^3}{7+t^4}$. Find the position of the object given that at $t = 1$, it is at the point 2. The position is $2 + \int_1^t \frac{3s^2}{7+s^4} ds$
4. An object moves on the x axis having velocity equal to $t \sin(2t^2)$. Find the position of the object given that at $t = 1$, it is at the point 1.
5. An object moves on the x axis having velocity equal to $\sec(t)$. Find the position of the object given that at $t = 1$, it is at the point -2 .
6. Find the indicated antiderivatives.
- (a) $\int \sec(3x) dx$ (d) $\int \frac{1}{\sqrt{5-4x^2}} dx$
- (b) $\int \sec^2(3x) \tan(3x) dx$ (e) $\int \frac{3}{x\sqrt{4x^2-5}} dx$
- (c) $\int \frac{1}{3+5x^2} dx$
7. Find the indicated antiderivatives.
- (a) $\int x \cosh(x^2 + 1) dx$ (d) $\int x \sin(x^2) dx$
- (b) $\int x^3 5^{x^4} dx$ (e) $\int x^5 \sqrt{2x^2 + 1} dx$ **Hint:** Let $u = 2x^2 + 1$.
- (c) $\int \sin(x) 7^{\cos(x)} dx$
8. Find $\int \sin^2(x) dx$. **Hint:** Derive and use $\sin^2(x) = \frac{1 - \cos(2x)}{2}$.
9. Find the indicated antiderivatives.
- (a) $\int \frac{\ln x}{x} dx$ (e) $\int \frac{1}{x\sqrt{x^2-9}} dx$ **Hint:** Let $x = 3u$.
- (b) $\int \frac{x^3}{3+x^4} dx$ (f) $\int \frac{\ln(x^2)}{x} dx$
- (c) $\int \frac{1}{x^2+2x+2} dx$ **Hint:** Complete the square in the denominator and then let $u = x + 1$. Remember the arctan function. (g) Find $\int \frac{x^3}{\sqrt{6x^2+5}} dx$
- (d) $\int \frac{1}{\sqrt{4-x^2}} dx$ (h) Find $\int x \sqrt[3]{6x+4} dx$
10. Find the indicated antiderivatives.
- (a) $\int x\sqrt{2x+4} dx$ (e) $\int \frac{1}{\sqrt{1+4x^2}} dx$
- (b) $\int x\sqrt{3x+2} dx$ (f) $\int \frac{x}{\sqrt{(3x-1)}} dx$
- (c) $\int \frac{1}{\sqrt{36-25x^2}} dx$ (g) $\int \frac{x}{\sqrt{5x+1}} dx$
- (d) $\int \frac{1}{\sqrt{9-4x^2}} dx$

(h) $\int \frac{1}{x\sqrt{9x^2-4}} dx$

(i) $\int \frac{1}{\sqrt{9+4x^2}} dx$

11. Find $\int \frac{1}{x^{1/3}+x^{1/2}} dx$. **Hint:** Try letting $x = u^6$ and use long division.
12. Suppose f is a function defined on \mathbb{R} and it satisfies the functional equation given by $f(a+b) = f(a) + f(b)$. Suppose also $f'(0) = k$. Find $f(x)$.
13. Suppose f is a function defined on \mathbb{R} having values in $(0, \infty)$ and it satisfies the functional equation $f(a+b) = f(a)f(b)$. Suppose also $f'(0) = k$. Find $f(x)$.
14. Suppose f is a function defined on $(0, \infty)$ having values in \mathbb{R} and it satisfies the functional equation $f(ab) = f(a) + f(b)$. Suppose also $f'(1) = k$. Find $f(x)$.
15. Suppose f is a function defined on \mathbb{R} and it satisfies the functional equation

$$f(a+b) = f(a) + f(b) + 3ab.$$

Suppose also that $\lim_{h \rightarrow 0} \frac{f(h)}{h} = 7$. Find $f(x)$ if possible.

8.3 Integration by Parts

Another technique for finding antiderivatives is called integration by parts and is based on the product rule. Recall the product rule. If u' and v' exist, then

$$(uv)'(x) = u'(x)v(x) + u(x)v'(x). \quad (8.2)$$

Therefore,

$$(uv)'(x) - u'(x)v(x) = u(x)v'(x)$$

Proposition 8.3.1 *Let u and v be differentiable functions for which*

$$\int u(x)v'(x) dx, \int u'(x)v(x) dx$$

are nonempty. Then

$$uv - \int u'(x)v(x) dx = \int u(x)v'(x) dx. \quad (8.3)$$

Proof: Let $F \in \int u'(x)v(x) dx$. Then

$$(uv - F)' = (uv)' - F' = (uv)' - u'v = uv'$$

by the chain rule. Therefore every function from the left in 8.3 is a function found in the right side of 8.3. Now let $G \in \int u(x)v'(x) dx$. Then $(uv - G)' = -uv' + (uv)' = u'v$ by the product rule. It follows that $uv - G \in \int u'(x)v(x) dx$ and so $G \in uv - \int u'(x)v(x) dx$. Thus every function from the right in 8.3 is a function from the left. ■

Example 8.3.2 *Find $\int x \sin(x) dx$.*

Let $u(x) = x$ and $v'(x) = \sin(x)$. Then applying 8.3,

$$\int x \sin(x) dx = (-\cos(x))x - \int (-\cos(x)) dx = -x \cos(x) + \sin(x) + C.$$

Example 8.3.3 Find $\int x \ln(x) dx$.

Let $u(x) = \ln(x)$ and $v'(x) = x$. Then from 8.3,

$$\begin{aligned} \int x \ln(x) dx &= \frac{x^2}{2} \ln(x) - \int \frac{x^2}{2} \left(\frac{1}{x} \right) \\ &= \frac{x^2}{2} \ln(x) - \int \frac{x}{2} = \frac{x^2}{2} \ln(x) - \frac{1}{4} x^2 + C \end{aligned}$$

The next example uses a trick.

Example 8.3.4 Find $\int \arctan(x) dx$.

Let $u(x) = \arctan(x)$ and $v'(x) = 1$. Then from 8.3,

$$\begin{aligned} \int \arctan(x) dx &= x \arctan(x) - \int x \left(\frac{1}{1+x^2} \right) dx \\ &= x \arctan(x) - \frac{1}{2} \int \frac{2x}{1+x^2} dx = x \arctan(x) - \frac{1}{2} \ln(1+x^2) + C. \end{aligned}$$

This trick works for \arctan , \ln , and various other inverse trig. functions.

Sometimes you want to find antiderivatives for something like $\int f g dx$ where $f^{(m)} = 0$ for some positive integer m . For example, $\int x^5 \sin x dx$. If you do integration by parts repeatedly, what do you get? Let $G'_1 = g, G'_2 = G_1, G'_3 = G_2$ etc. Then the first application of integration by parts yields $f G_1 - \int G_1 f' dx$. The next application of integration by parts yields $f G_1 - G_2 f' + \int G_2 f'' dx$. Yet another application of integration by parts yields $f G_1 - G_2 f' + G_3 f'' - \int G_3 f''' dx$. Eventually the process will stop because a high enough derivative of f equals zero. This justifies the following procedure for finding antiderivatives in this case.

Procedure 8.3.5 Suppose $f^{(m)} = 0$ for some m a positive integer and let $G'_k = G_{k-1}$ for all k and $G_0 = g$. Then

$$\int f g dx = f G_1 - f' G_2 + f'' G_3 - f''' G_4 + \cdots$$

Just keep writing these terms, alternating signs until the process yields a zero. Then add on an arbitrary constant of integration and stop. Sometimes people remember this in the form of a table.

		g
f	$\xrightarrow{+}$	G_1
f'	$\xrightarrow{-}$	G_2
f''	$\xrightarrow{+}$	G_3
f'''	$\xrightarrow{-}$	G_4

Thus you fill in the table until the left column ends in a 0 and then do the arrows, $f G_1 - f' G_2 + f'' G_3 \cdots$ till the process ends. Then add C , a constant of integration.

Example 8.3.6 Find $\int x^5 \sin x dx$.

From the above procedure, and letting $f(x) = x^5$, this equals

$$\begin{aligned} & x^5(-\cos(x)) - 5x^4(-\sin(x)) + 20x^3(\cos(x)) - 60x^2(\sin(x)) \\ & + 120x(-\cos(x)) - 120(-\sin(x)) + C. \end{aligned}$$

To determine the distance an object moves for $t \in [a, b]$, one computes

$$\int_a^b |v(t)| dt$$

The reason is as follows. If $v(t)$ is nonnegative on an interval $[c, d]$, this means the position is increasing. To find the distance travelled, you would consider $r(d) - r(c) = \int_c^d v(t) dt$. If $v(t) < 0$, on an interval $[c, d]$ this means $r(t)$ is decreasing. Thus the distance on this interval is $r(c) - r(d)$ which equals $\int_c^d -v(t) dt = \int_c^d |v(t)| dt$. Splitting the interval into sub-intervals on which the velocity is either positive or negative, one obtains that the distance traveled is $\int_a^b |v(t)| dt$. This motivates the following definition.

Definition 8.3.7 Let the position of an object moving on \mathbb{R} be denoted as $r(t)$. Then the distance moved for $t \in [a, b]$ is

$$\int_a^b |r'(t)| dt$$

Example 8.3.8 Suppose the velocity is $v(t) = t - t^3$. Find the distance the object moves on the real line for $t \in [0, 2]$.

As just explained, it is $\int_0^2 |t - t^3| dt$. You must split this up into intervals on which you can remove the absolute values. $t - t^3 \geq 0$ on $[0, 1]$ and it is ≤ 0 on $[1, 2]$ so the total distance travelled is

$$\int_0^1 (t - t^3) dt + \int_1^2 (-t + t^3) dt = \frac{5}{2}$$

Sometimes people want to use a shortcut on problems like this. They want to say that an antiderivative is $\left| \frac{t^2}{2} - \frac{t^4}{4} \right|$ and then plug in the end points and evaluate. This is totally wrong because the function just described is not an antiderivative of the function $t \rightarrow |t - t^3|$!

8.4 Exercises

1. Find the following antiderivatives.

(a) $\int x^3 e^{-3x} dx$

(d) $\int x^6 \sin(2x) dx$

(b) $\int x^4 \cos x dx$

(e) $\int x^3 \cos(x^2) dx$

(c) $\int x^5 e^x dx$

2. Find the following antiderivatives.

(a) $\int x e^{-3x} dx$

(b) $\int \frac{1}{x(\ln(|x|))^2} dx$

(c) $\int x\sqrt{2-x}dx$

(d) $\int (\ln|x|)^2 dx$ **Hint:** Let $u(x) = (\ln|x|)^2$ and $v'(x) = 1$.

(e) $\int x^3 \cos(x^2) dx$

3. Show that $\int \sec^3(x) dx =$

$$\frac{1}{2} \tan(x) \sec(x) + \frac{1}{2} \ln|\sec x + \tan x| + C.$$

4. Find $\int \frac{xe^x}{(1+x)^2} dx$.

5. Consider the following argument. Integrate by parts, letting $u(x) = x$ and $v'(x) = \frac{1}{x^2}$ to get

$$\int \frac{1}{x} dx = \int x \left(\frac{1}{x^2} \right) dx = \left(-\frac{1}{x} \right) x + \int \frac{1}{x} dx = -1 + \int \frac{1}{x} dx.$$

Now subtracting $\int \frac{1}{x} dx$ from both sides, $0 = -1$. Is there anything wrong here? If so, what?

6. Find the following antiderivatives.

(a) $\int x^3 \arctan(x) dx$

(e) $\int x \arcsin(x) dx$

(b) $\int x^3 \ln(x) dx$

(f) $\int \cos(2x) \sin(3x) dx$

(c) $\int x^2 \sin(x) dx$

(g) $\int x^3 e^{x^2} dx$

(d) $\int x^2 \cos(x) dx$

(h) $\int x^3 \cos(x^2) dx$

7. Find the antiderivatives

(a) $\int x^2 \sin x dx$

(e) $\int (x+2)^2 e^x dx$

(b) $\int x^3 \sin x dx$

(f) $\int x^3 2^x dx$

(c) $\int x^3 7^x dx$

(g) $\int \sec^3(2x) \tan(2x) dx$

(d) $\int x^2 \ln x dx$

(h) $\int x^2 7^x dx$

8. Solve the initial value problem $y'(x) = f(x)$, $\lim_{x \rightarrow 0^+} y(x) = 1$ where $f(x)$ is each of the integrands in Problem 7.

9. Solve the initial value problem $y'(x) = f(x)$, $\lim_{x \rightarrow 0^+} y(x) = 2$ where $f(x)$ is each of the integrands in Problem 6.

10. Try doing $\int \sin^2 x dx$ the obvious way. If you do not make any mistakes, the process will go in circles. Now do it by taking

$$\int \sin^2 x dx = x \sin^2 x - 2 \int x \sin x \cos x dx = x \sin^2 x - \int x \sin(2x) dx.$$

11. An object moves on the x axis having velocity equal to $t \sin t$. Find the position of the object given that at $t = 1$, it is at the point 2.

12. An object moves on the x axis having velocity equal to $\sec^3(t)$. Find the position of the object given that at $t = 0$, it is at the point 2. **Hint:** You might want to use Problem 3.
13. Find the antiderivatives.
- | | |
|--------------------------------------|---------------------------------|
| (a) $\int x \cos(x^2) dx$ | (e) $\int \arcsin(x) dx$ |
| (b) $\int \sin(\sqrt{x}) dx$ | (f) $\int \sec^3(x) \tan(x) dx$ |
| (c) $\int \ln(\sin(x)) \cos(x) dx$ | (g) $\int \tan^2(x) \sec(x) dx$ |
| (d) $\int \cos^4(x) dx$ | |
14. A car is moving at 14 feet per second when the driver applies the brake causing the car to slow down at the constant rate of 2 feet per second per second until it stops. How far does the car travel during the time the brake was applied?
15. Suppose you have the graphs of two functions $y = f(x)$ and $y = g(x)$ defined for $x \in [a, b]$. How would you define the area between the two graphs for $x \in [a, b]$? You would first consider an approximation by considering little rectangles of height $|f(z_i) - g(z_i)|$ and width $x_i - x_{i-1}$ where $a = x_0 < \dots < x_n = b$ and $z_i \in [x_{i-1}, x_i]$ and adding the areas of these. It is reasonable to suppose that as the norm of the partition becomes increasingly small so that the rectangles get increasingly thin that what occurs in the limit should be the **definition** of the area between the two graphs. But this limit is defined as $\int_a^b |f(x) - g(x)| dx$. Find the area between the two given graphs on the given interval.
- | |
|---|
| (a) $f(x) = x, g(x) = x - x^3, x \in [0, 3]$ |
| (b) $f(x) = \sin(x), g(x) = \cos(x), x \in [0, 2\pi]$ |
| (c) $f(x) = e^x, g(x) = \ln(x), x \in [1, 2]$ |

8.5 Trig. Substitutions

Certain antiderivatives are easily obtained by making an auspicious substitution involving a trig. function. The technique will be illustrated by presenting examples.

Example 8.5.1 Find $\int \frac{1}{(x^2+2x+2)^2} dx$.

Complete the square as before and write

$$\int \frac{1}{(x^2+2x+2)^2} dx = \int \frac{1}{((x+1)^2+1)^2} dx$$

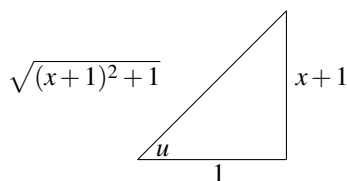
Use the following substitution next.

$$x+1 = \tan u \tag{8.4}$$

so $dx = (\sec^2 u) du$. Therefore, this last indefinite integral becomes

$$\begin{aligned} \int \frac{\sec^2 u}{(\tan^2 u + 1)^2} du &= \int (\cos^2 u) du = \int \frac{1 + \cos 2u}{2} du \\ &= \frac{u}{2} + \frac{\sin 2u}{4} + C = \frac{u}{2} + \frac{2 \sin u \cos u}{4} + C \end{aligned}$$

Next write this in terms of x using the following device based on the following picture.



In this picture which is descriptive of 8.4, $\sin u = \frac{x+1}{\sqrt{(x+1)^2 + 1}}$ and $\cos u = \frac{1}{\sqrt{(x+1)^2 + 1}}$. Therefore, putting in this information to change back to the x variable,

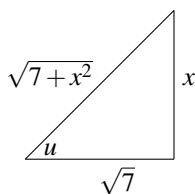
$$\begin{aligned} &\int \frac{1}{(x^2 + 2x + 2)^2} dx \\ &= \frac{1}{2} \arctan(x+1) + \frac{1}{2} \frac{x+1}{\sqrt{(x+1)^2 + 1}} \frac{1}{\sqrt{(x+1)^2 + 1}} + C \\ &= \frac{1}{2} \arctan(x+1) + \frac{1}{2} \frac{x+1}{(x+1)^2 + 1} + C. \end{aligned}$$

Example 8.5.2 Find $\int \frac{1}{\sqrt{x^2+7}} dx$.

Let $x = \sqrt{7} \tan u$ so $dx = \sqrt{7} (\sec^2 u) du$. Making the substitution, consider

$$\int \frac{1}{\sqrt{7} \sqrt{\tan^2 u + 1}} \sqrt{7} (\sec^2 u) du = \int (\sec u) du = \ln |\sec u + \tan u| + C$$

Now the following diagram is descriptive of the above transformation.



Using the above diagram, $\sec u = \frac{\sqrt{7+x^2}}{\sqrt{7}}$ and $\tan u = \frac{x}{\sqrt{7}}$. Therefore, restoring the x variable,

$$\int \frac{1}{\sqrt{x^2+7}} dx = \ln \left| \frac{\sqrt{7+x^2}}{\sqrt{7}} + \frac{x}{\sqrt{7}} \right| + C = \ln |\sqrt{7+x^2} + x| + C.$$

Note the constant C changed in going from the top to the bottom line. It is $C - \ln \sqrt{7}$ but it is customary to simply write this as C because C is arbitrary.

Example 8.5.3 Find $\int (4x^2 + 3)^{1/2} dx$.

Let $2x = \sqrt{3} \tan u$ so $2dx = \sqrt{3} \sec^2(u) du$. Then making the substitution,

$$\sqrt{3} \int (\tan^2 u + 1)^{1/2} \frac{\sqrt{3}}{2} \sec^2(u) du = \frac{3}{2} \int \sec^3(u) du. \quad (8.5)$$

Now use integration by parts to obtain

$$\begin{aligned} \int \sec^3(u) du &= \int \sec^2(u) \sec(u) du \\ &= \tan(u) \sec(u) - \int \tan^2(u) \sec(u) du \\ &= \tan(u) \sec(u) - \int (\sec^2(u) - 1) \sec(u) du \\ &= \tan(u) \sec(u) + \int \sec(u) du - \int \sec^3(u) du \\ &= \tan(u) \sec(u) + \ln |\sec(u) + \tan(u)| - \int \sec^3(u) du \end{aligned}$$

Therefore,

$$2 \int \sec^3(u) du = \tan(u) \sec(u) + \ln |\sec(u) + \tan(u)| + C$$

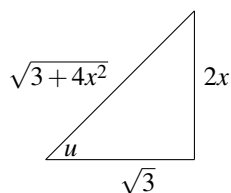
and so

$$\int \sec^3(u) du = \frac{1}{2} [\tan(u) \sec(u) + \ln |\sec(u) + \tan(u)|] + C. \quad (8.6)$$

Now it follows from 8.5 that in terms of u the set of antiderivatives is given by

$$\frac{3}{4} [\tan(u) \sec(u) + \ln |\sec(u) + \tan(u)|] + C$$

Use the following diagram to change back to the variable x .



From the diagram, $\tan(u) = \frac{2x}{\sqrt{3}}$ and $\sec(u) = \frac{\sqrt{3+4x^2}}{\sqrt{3}}$. Therefore,

$$\begin{aligned} &\int (4x^2 + 3)^{1/2} dx \\ &= \frac{3}{4} \left[\frac{2x}{\sqrt{3}} \frac{\sqrt{3+4x^2}}{\sqrt{3}} + \ln \left| \frac{\sqrt{3+4x^2}}{\sqrt{3}} + \frac{2x}{\sqrt{3}} \right| \right] + C \\ &= \frac{3}{4} \left[\frac{2x}{\sqrt{3}} \frac{\sqrt{3+4x^2}}{\sqrt{3}} + \ln \left| \frac{\sqrt{3+4x^2}}{\sqrt{3}} + \frac{2x}{\sqrt{3}} \right| \right] + C \\ &= \frac{1}{2} x \sqrt{3+4x^2} + \frac{3}{4} \ln |\sqrt{3+4x^2} + 2x| + C \end{aligned}$$

Note that these examples involved something of the form $(a^2 + (bx)^2)$ and the trig substitution $bx = a \tan u$ was the right one to use. This is the auspicious substitution which often simplifies these sorts of problems. However, there is a possibly better way to do these kinds.

Example 8.5.4 Find $\int (4x^2 + 3)^{1/2} dx$ another way.

Let $2x = \sqrt{3} \sinh u$ and so $2dx = \sqrt{3} \cosh(u) du$. Then substituting in the integral leads to

$$\begin{aligned} & \int \sqrt{3} \sqrt{1 + \sinh^2(u)} \frac{\sqrt{3}}{2} \cosh(u) du = \frac{3}{2} \int \cosh^2(u) du + C \\ &= \frac{3}{4} \cosh(u) \sinh(u) + \frac{3}{4} u + C = \frac{3}{4} \sqrt{1 + \sinh^2(u)} \sinh(u) + \frac{3}{4} u + C \\ &= \frac{1}{2} x \sqrt{3 + 4x^2} + \frac{3}{4} \sinh^{-1} \left(\frac{2x}{\sqrt{3}} \right) + C \end{aligned}$$

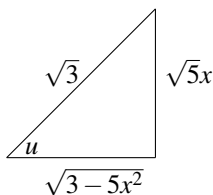
This other way is often used by computer algebra systems. If you solve for $\sinh^{-1} x$ in terms of \ln , you get the same set of antiderivatives. The function \sinh^{-1} is also written as $\operatorname{arcsinh}$ by analogy to the trig. functions also as asinh .

Example 8.5.5 Find $\int \sqrt{3 - 5x^2} dx$.

In this example, let $\sqrt{5}x = \sqrt{3} \sin(u)$ so $\sqrt{5}dx = \sqrt{3} \cos(u) du$. The reason this might be a good idea is that it will get rid of the square root sign as shown below. Making the substitution, leads to

$$\frac{3}{2\sqrt{5}} u + \frac{3}{\sqrt{5}} \sin(2u) + C = \frac{3}{2\sqrt{5}} u + \frac{3}{2\sqrt{5}} \sin u \cos u + C$$

The appropriate diagram is the following.



From the diagram, $\sin(u) = \frac{\sqrt{5}x}{\sqrt{3}}$ and $\cos(u) = \frac{\sqrt{3 - 5x^2}}{\sqrt{3}}$. Therefore, changing back to x ,

$$\begin{aligned} & \int \sqrt{3 - 5x^2} dx = \\ &= \frac{3}{10} \sqrt{5} \arcsin \left(\frac{1}{3} \sqrt{15} x \right) + \frac{1}{2} x \sqrt{3 - 5x^2} + C \end{aligned}$$

Example 8.5.6 Find $\int \sqrt{5x^2 - 3} dx$.

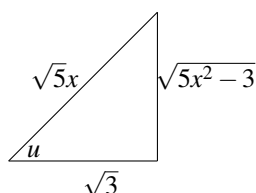
In this example, let $\sqrt{5}x = \sqrt{3}\sec(u)$ so $\sqrt{5}dx = \sqrt{3}\sec(u)\tan(u)du$. Then changing the variable, consider

$$\begin{aligned} & \sqrt{3} \int \sqrt{\sec^2(u) - 1} \frac{\sqrt{3}}{\sqrt{5}} \sec(u) \tan(u) du = \frac{3}{\sqrt{5}} \int \tan^2(u) \sec(u) du \\ &= \frac{3}{\sqrt{5}} \left[\int \sec^3(u) du - \int \sec(u) du \right] \end{aligned}$$

Now from 8.6, this equals

$$\begin{aligned} & \frac{3}{\sqrt{5}} \left[\frac{1}{2} [\tan(u) \sec(u) + \ln|\sec(u) + \tan(u)|] - \ln|\tan(u) + \sec(u)| \right] + C \\ &= \frac{3}{2\sqrt{5}} \tan(u) \sec(u) - \frac{3}{2\sqrt{5}} \ln|\sec(u) + \tan(u)| + C. \end{aligned}$$

Now it is necessary to change back to x . The diagram is as follows.



Therefore, $\tan(u) = \frac{\sqrt{5x^2-3}}{\sqrt{3}}$ and $\sec(u) = \frac{\sqrt{5}x}{\sqrt{3}}$ and so

$$\begin{aligned} & \int \sqrt{5x^2 - 3} dx \\ &= \frac{1}{2} \left(\sqrt{5x^2 - 3} \right) x - \frac{3}{10} \sqrt{5} \ln \left| \sqrt{5}x + \sqrt{(-3 + 5x^2)} \right| + C \end{aligned}$$

To summarize, here is a short table of auspicious substitutions corresponding to certain expressions.

Table Of Auspicious Substitutions

Expression	$a^2 + b^2x^2$	$a^2 - b^2x^2$	$a^2x^2 - b^2$
Trig. substitution	$bx = a \tan(u)$	$bx = a \sin(u)$	$ax = b \sec(u)$
Hyperbolic substitution	$bx = a \sinh(u)$		

Of course there are no “magic bullets” but these substitutions will often simplify an expression enough to allow you to find an antiderivative. These substitutions are often especially useful when the expression is enclosed in a square root.

8.6 Exercises

1. Find the antiderivatives.

- | | |
|---|----------------------------------|
| (a) $\int \frac{x}{\sqrt{4-x^2}} dx$ | (f) $\int (\sqrt{9-16x^2})^3 dx$ |
| (b) $\int \frac{3}{\sqrt{36-25x^2}} dx$ | (g) $\int (\sqrt{16-x^2})^5 dx$ |
| (c) $\int \frac{3}{\sqrt{16-25x^2}} dx$ | (h) $\int \sqrt{25-36x^2} dx$ |
| (d) $\int \frac{1}{\sqrt{4-9x^2}} dx$ | (i) $\int (\sqrt{4-9x^2})^3 dx$ |
| (e) $\int \frac{1}{\sqrt{36-x^2}} dx$ | (j) $\int \sqrt{1-9x^2} dx$ |

2. Find the antiderivatives.

- | | |
|-------------------------------|----------------------------------|
| (a) $\int \sqrt{36x^2-25} dx$ | (c) $\int (\sqrt{16x^2-9})^3 dx$ |
| (b) $\int \sqrt{x^2-4} dx$ | (d) $\int \sqrt{25x^2-16} dx$ |

3. Find the antiderivatives.

- | | |
|---|--|
| (a) $\int \frac{1}{26+x^2-2x} dx$ Hint: Complete the square. | (f) $\int \frac{1}{(16+25(x-3)^2)^2} dx$ |
| (b) $\int \sqrt{x^2+9} dx$ | (g) $\int \frac{1}{261+25x^2-150x} dx$ Hint: Complete the square. |
| (c) $\int \sqrt{4x^2+25} dx$ | (h) $\int (\sqrt{25x^2+9})^3 dx$ |
| (d) $\int x\sqrt{4x^4+9} dx$ | (i) $\int \frac{1}{25+16x^2} dx$ |
| (e) $\int x^3\sqrt{4x^4+9} dx$ | |

4. Find the antiderivatives. **Hint:** Complete the square.

- | | |
|---|--|
| (a) $\int \sqrt{4x^2+16x+15} dx$ | (d) $\int \frac{3}{\sqrt{-5-x^2-6x}} dx$ |
| (b) $\int \sqrt{x^2+6x} dx$ | (e) $\int \frac{1}{\sqrt{9-16x^2-32x}} dx$ |
| (c) $\int \frac{3}{\sqrt{-32-9x^2-36x}} dx$ | (f) $\int \sqrt{4x^2+16x+7} dx$ |

5. Find $\int x^5 \sqrt{1+x^4} dx$.

6. Find $\int \frac{x}{\sqrt{1-x^4}} dx$. **Hint:** Try $x^2 = \sin(u)$.

8.7 Partial Fractions

The main technique for finding antiderivatives in the case $f(x) = \frac{p(x)}{q(x)}$ for p and q polynomials is the technique of partial fractions. Before presenting this technique, a few more examples are presented. These examples are typical of the kind of thing you end up doing after you have found the partial fractions expansion.

Example 8.7.1 Find $\int \frac{1}{x^2+2x+2} dx$.

To do this, complete the square in the denominator to write

$$\int \frac{1}{x^2 + 2x + 2} dx = \int \frac{1}{(x+1)^2 + 1} dx$$

Now change the variable, letting $u = x + 1$, so that $du = dx$. Then the last indefinite integral reduces to

$$\int \frac{1}{x^2 + 2x + 2} dx = \arctan(x+1) + C.$$

Example 8.7.2 Find $\int \frac{1}{3x+5} dx$.

Let $u = 3x + 5$ so $du = 3dx$. Then you obtain $\int \frac{1}{3x+5} dx = \frac{1}{3} \ln|3x+5| + C$.

Example 8.7.3 Find $\int \frac{3x+2}{x^2+x+1} dx$.

First complete the square in the denominator.

$$\int \frac{3x+2}{x^2+x+1} dx = \int \frac{3x+2}{x^2+x+\frac{1}{4}+\frac{3}{4}} dx = \int \frac{3x+2}{(x+\frac{1}{2})^2+\frac{3}{4}} dx.$$

Now let $(x+\frac{1}{2})^2 = \frac{3}{4}u^2$ so that $x+\frac{1}{2} = \frac{\sqrt{3}}{2}u$. Therefore, $dx = \frac{\sqrt{3}}{2}du$ and changing the variable, one obtains

$$\begin{aligned} &= \frac{\sqrt{3}}{2} \left(2\sqrt{3} \int \frac{u}{u^2+1} du - \frac{2}{3} \int \frac{1}{u^2+1} du \right) \\ &= \frac{3}{2} \ln(u^2+1) - \frac{\sqrt{3}}{3} \arctan u + C \end{aligned}$$

Therefore, $\int \frac{3x+2}{x^2+x+1} dx =$

$$\frac{3}{2} \ln \left(\left(\frac{2}{\sqrt{3}} \left(x + \frac{1}{2} \right) \right)^2 + 1 \right) - \frac{\sqrt{3}}{3} \arctan \left(\frac{2}{\sqrt{3}} \left(x + \frac{1}{2} \right) \right) + C.$$

The method of partial fractions splits rational functions into a sum of functions which are like those which were just done successfully. In using this method it is essential that in the rational function the degree of the numerator is smaller than the degree of the denominator. Lemma 1.12.3 on dividing polynomials implies the following important corollary.

Corollary 8.7.4 Let $f(x)$ and $g(x)$ be polynomials. Then there exists a polynomial, $r(x)$ such that the degree of $r(x) < \text{degree of } g(x)$ and a polynomial, $q(x)$ such that

$$\frac{f(x)}{g(x)} = q(x) + \frac{r(x)}{g(x)}.$$

Here is an example where the degree of the numerator exceeds the degree of the denominator.

Example 8.7.5 Find $\int \frac{3x^5+7}{x^2-1} dx$.

In this case the degree of the numerator is larger than the degree of the denominator and so long division must first be used. Thus

$$\frac{3x^5 + 7}{x^2 - 1} = 3x^3 + 3x + \frac{7 + 3x}{x^2 - 1}$$

Recall the process of long division from elementary school. The only difference is that you use x raised to powers rather than 10 raised to powers. I am reviewing the algorithm in what follows. The first term on the top is $3x^3$ because $3x^3$ times x^2 gives $3x^5$ which will cancel the first term of the $3x^5 + 0x^4 + 0x^3 + 0x^2 + 0x + 7$. Then you multiply the $x^2 + 0x - 1$ by the $3x^3$ and subtract. Then you do the same process on $3x^3 + 0x^2 + 0x$. A more careful presentation of this algorithm is in my pre calculus book published by worldwide center of math.

$$\begin{array}{r} x^2 + 0x - 1 \quad \overline{) 3x^5 + 0x^4 + 0x^3 + 0x^2 + 0x + 7} \\ \underline{3x^5 + 0x^4 - 3x^3} \\ 3x^3 + 0x^2 + 0x \\ \underline{3x^3 + 0x^2 - 3x} \\ 3x + 7 \end{array}$$

Now look for a partial fractions expansion of the form

$$\frac{7 + 3x}{x^2 - 1} = \frac{a}{(x - 1)} + \frac{b}{(x + 1)}.$$

Therefore, $7 + 3x = a(x + 1) + b(x - 1)$. Letting $x = 1$, $a = 5$. Then letting $x = -1$, it follows $b = -2$. Therefore,

$$\frac{7 + 3x}{x^2 - 1} = \frac{5}{x - 1} - \frac{2}{x + 1}$$

and so $\frac{3x^5 + 7}{x^2 - 1} = 3x^3 + 3x + \frac{5}{x - 1} - \frac{2}{x + 1}$. Therefore,

$$\int \frac{3x^5 + 7}{x^2 - 1} dx = \frac{3}{4}x^4 + \frac{3}{2}x^2 + 5 \ln(x - 1) - 2 \ln(x + 1) + C.$$

Here is another example.

Example 8.7.6 Find $\int \frac{7x^3 + 19x^2 + 20x + 8}{(2x + 1)(x + 3)(x^2 + x + 1)} dx$.

The degree of the top is less than the degree of the bottom and so we look for a partial fractions expansion of the form

$$\frac{7x^3 + 19x^2 + 20x + 8}{(2x + 1)(x + 3)(x^2 + x + 1)} = \frac{a}{2x + 1} + \frac{b}{x + 3} + \frac{cx + d}{x^2 + x + 1}$$

The reason the last term has a $cx + d$ on the top is that the bottom of the fraction is an irreducible polynomial. Now it is just a matter of finding a, b, c, d . Multiply both sides by $(x + 3)$ and then plug in $x = -3$.

$$\frac{7(-3)^3 + 19(-3)^2 + 20(-3) + 8}{(2(-3) + 1)((-3)^2 + (-3) + 1)} = 2 = b$$

Next multiply both sides by $(2x+1)$ and plug in $x = -1/2$.

$$\frac{7(-1/2)^3 + 19(-1/2)^2 + 20(-1/2) + 8}{((-1/2) + 3)((-1/2)^2 + (-1/2) + 1)} = 1 = a$$

Plug these values in on the right and subtract from both sides.

$$\begin{aligned} \frac{7x^3 + 19x^2 + 20x + 8}{(2x+1)(x+3)(x^2+x+1)} - \left(\frac{1}{2x+1} + \frac{2}{x+3} \right) &= \frac{cx+d}{x^2+x+1} \\ \frac{x+1}{x^2+x+1} &= \frac{cx+d}{x^2+x+1} \end{aligned}$$

Now it is obvious that $c = 1$ and $d = 1$.

$$\begin{aligned} \int \frac{7x^3 + 19x^2 + 20x + 8}{(2x+1)(x+3)(x^2+x+1)} dx &= \int \left(\frac{1}{2x+1} + \frac{2}{x+3} + \frac{x+1}{x^2+x+1} \right) dx \\ &= \frac{1}{2} \ln(x^2+x+1) + 2 \ln(x+3) + \frac{1}{2} \ln\left(x + \frac{1}{2}\right) - \frac{1}{6} \sqrt{3} \pi \\ &\quad + \frac{1}{3} \sqrt{3} \arctan \sqrt{3} \left(\frac{2}{3}x + \frac{1}{3} \right) + C \end{aligned}$$

What is done when the factors are repeated?

Example 8.7.7 Find $\int \frac{3x+7}{(x+2)^2(x+3)} dx$.

First observe that the degree of the numerator is less than the degree of the denominator. In this case the correct form of the partial fraction expansion is

$$\frac{a}{(x+2)} + \frac{b}{(x+2)^2} + \frac{c}{(x+3)}.$$

The reason there are two terms devoted to $(x+2)$ is that this is squared. Computing the constants yields

$$\frac{3x+7}{(x+2)^2(x+3)} = \frac{1}{(x+2)^2} + \frac{2}{x+2} - \frac{2}{x+3}$$

and therefore,

$$\int \frac{3x+7}{(x+2)^2(x+3)} dx = -\frac{1}{x+2} + 2 \ln|x+2| - 2 \ln|x+3| + C.$$

Example 8.7.8 Find the proper form for the partial fractions expansion of

$$\frac{x^3 + 7x + 9}{(x^2 + 2x + 2)^3 (x+2)^2 (x+1) (x^2 + 1)}.$$

First check to see if the degree of the numerator is smaller than the degree of the denominator. Since this is the case, look for a partial fractions decomposition in the following form.

$$\frac{ax+b}{(x^2+2x+2)} + \frac{cx+d}{(x^2+2x+2)^2} + \frac{ex+f}{(x^2+2x+2)^3} +$$

$$\frac{A}{(x+2)} + \frac{B}{(x+2)^2} + \frac{D}{(x+1)} + \frac{gx+h}{x^2+1}.$$

These examples illustrate what to do when using the method of partial fractions. You first check to be sure the degree of the numerator is less than the degree of the denominator. If this is not so, do a long division. Then you factor the denominator into a product of factors, some linear of the form $ax+b$ and others quadratic, ax^2+bx+c which cannot be factored further. Next follow the procedure illustrated in the above examples and summarized below.

Warning: When you use partial fractions, **be sure you look for something which is of the right form.** Otherwise you may be looking for something which is not there. The rules are summarized next.

Rules For Finding Partial Fractions Expansion Of A Rational Function

1. Check to see if the numerator has smaller degree than the denominator. If this is not so, correct the situation by doing long division.
2. Factor the denominator into a product of linear factors, (Things like $(ax+b)$) and irreducible quadratic factors, (Things like (ax^2+bx+c) where $b^2-4ac < 0$.)¹
3. Let m, n be positive integers. Corresponding to $(ax+b)^m$ in the denominator, you should have a sum of the form $\sum_{i=1}^m \frac{c_i}{(ax+b)^i}$ in the partial fractions expansion. Here the c_i are the constants to be found. Corresponding to $(ax^2+bx+c)^n$ in the denominator where $b^2-4ac < 0$, you should have a sum of the form $\sum_{i=1}^n \frac{p_ix+q_i}{(ax^2+bx+c)^i}$ in the partial fractions expansion. Here the p_i and q_i are to be found.
4. Find the constants, c_i , p_i , and q_i . Use whatever method you like. You might see if you can make up new ways to do this if you like. If you have followed steps 1 - 3 correctly, it will work out. However, be sure to search for something which is actually there. Otherwise, you won't find it.

The above technique for finding the coefficients is fine but some people like to do it other ways. It really does not matter how you do it. Here is another example.

Example 8.7.9 Find the partial fractions expansion for

$$\frac{15x^4 + 44x^3 + 71x^2 + 64x + 28 + 2x^5}{(x+2)^2(x^2+2x+2)^2}$$

¹Of course this factoring of the denominator is easier said than done. In general you cannot do it at all. Of course there are big theorems which guarantee the existence of such a factorization but these theorems do not tell how to find it. This is an example of the gap between theory and practice which permeates mathematics.

The degree of the top is 4 and the degree of the bottom is 6 so you do not need to do long division. You do have to look for the right thing however. The correct form for the partial fractions expansion is

$$\begin{aligned} & \frac{a}{x+2} + \frac{b}{(x+2)^2} + \frac{cx+d}{x^2+2x+2} + \frac{ex+f}{(x^2+2x+2)^2} \\ = & \frac{15x^4 + 44x^3 + 71x^2 + 64x + 28 + 2x^5}{(x+2)^2(x^2+2x+2)^2} \end{aligned}$$

Multiply both sides by $(x+2)^2$ and then plug in $x = -2$.

$$b = \frac{15(-2)^4 + 44(-2)^3 + 71(-2)^2 + 64(-2) + 28 + 2(-2)^5}{((-2)^2 + 2(-2) + 2)^2} = 2$$

Now subtract the term involving b from both sides.

$$\begin{aligned} & \frac{a}{x+2} + \frac{cx+d}{x^2+2x+2} + \frac{ex+f}{(x^2+2x+2)^2} = \\ & \frac{15x^4 + 44x^3 + 71x^2 + 64x + 28 + 2x^5}{(x+2)^2(x^2+2x+2)^2} - \frac{2}{(x+2)^2} \\ = & \frac{1}{(x+2)(x^2+2x+2)^2} (2x^4 + 9x^3 + 18x^2 + 19x + 10) \end{aligned}$$

Multiply both sides by $x+2$ and plug in $x = -2$.

$$a = \frac{1}{((-2)^2 + 2(-2) + 2)^2} (2(-2)^4 + 9(-2)^3 + 18(-2)^2 + 19(-2) + 10) = 1$$

Subtract this term involving a from both sides.

$$\begin{aligned} & \frac{cx+d}{x^2+2x+2} + \frac{ex+f}{(x^2+2x+2)^2} \\ = & \frac{(2x^4 + 9x^3 + 18x^2 + 19x + 10)}{(x+2)(x^2+2x+2)^2} - \frac{1}{x+2} \\ = & \frac{x^3 + 3x^2 + 4x + 3}{(x^2+2x+2)^2} \end{aligned}$$

Add the fractions on the left.

$$\frac{cx^3 + (2c+d)x^2 + (2c+2d+e)x + (2d+f)}{(x^2+2x+2)^2} = \frac{x^3 + 3x^2 + 4x + 3}{(x^2+2x+2)^2}$$

Now you see $c = 1, d = 1, e = 0, f = 1$.

It follows the partial fractions expansion is

$$\frac{1}{x+2} + \frac{2}{(x+2)^2} + \frac{x+1}{x^2+2x+2} + \frac{1}{(x^2+2x+2)^2}.$$

One other thing should be mentioned. Suppose you wanted to find the integral in this example. The first three terms are by now routine. How about the last one?

Example 8.7.10 Find $\int \frac{1}{(x^2+2x+2)^2} dx$.

First complete the square to write this as $\int \frac{1}{((x+1)^2+1)^2} dx$. Now do a trig. substitution. You should let $x+1 = \tan \theta$. Then the integral becomes

$$\begin{aligned} \int \frac{1}{\sec^4 \theta} \sec^2 \theta d\theta &= \int \cos^2 \theta d\theta = \int \frac{1 + \cos(2\theta)}{2} d\theta \\ &= \frac{\theta}{2} + \frac{2 \sin \theta \cos \theta}{4} + C. \end{aligned}$$

Setting up a little triangle as in the section on trig. substitutions, you can restore the original variables to obtain

$$\frac{1}{2} \arctan(x+1) + \frac{1}{2} \left(\frac{x+1}{x^2+2x+2} \right) + C.$$

The theory of partial fractions is in Problem 40 on Page 48.

8.8 Rational Functions of Trig. Functions

There is a technique which reduces certain kinds of integrals involving trig. functions to the technique of partial fractions. This is illustrated in the following example.

Example 8.8.1 Find $\int \frac{\cos \theta}{1+\cos \theta} d\theta$.

The integrand is an example of a rational function of cosines and sines. When such a thing occurs there is a substitution which will reduce the integrand to a rational function like those above which can then be integrated using partial fractions. The substitution is $u = \tan\left(\frac{\theta}{2}\right)$. Thus in this example, $du = \left(1 + \tan^2\left(\frac{\theta}{2}\right)\right) \frac{1}{2} d\theta$ and so in terms of this new variable, the indefinite integral is

$$\int \frac{2 \cos(2 \arctan u)}{(1 + \cos(2 \arctan u))(1 + u^2)} du.$$

You can evaluate $\cos(2 \arctan u)$ exactly. This equals $2 \cos^2(\arctan u) - 1$. Setting up a little triangle as above, $\cos(\arctan u)$ equals $1/\sqrt{1+u^2}$ and so the integrand reduces to

$$\frac{2 \left(2 \left(1/\sqrt{1+u^2} \right)^2 - 1 \right)}{\left(1 + \left(2 \left(1/\sqrt{1+u^2} \right)^2 - 1 \right) \right) (1+u^2)} = \frac{1-u^2}{1+u^2} = -1 + \frac{2}{1+u^2}$$

therefore, in terms of u , the antiderivative equals $-u + 2 \arctan u$. Now replace u to obtain

$$-\tan\left(\frac{\theta}{2}\right) + 2 \arctan\left(\tan\left(\frac{\theta}{2}\right)\right) + C.$$

This procedure can be expected to work in general. Suppose you want to find

$$\int \frac{p(\cos \theta, \sin \theta)}{q(\cos \theta, \sin \theta)} d\theta$$

where p and q are polynomials in each argument. Make the substitution $u = \tan \frac{\theta}{2}$. As above this means

$$du = \left(1 + \tan^2 \left(\frac{\theta}{2}\right)\right) \frac{1}{2} d\theta = \frac{1}{2} (1 + u^2) d\theta.$$

It remains to substitute for $\sin \theta$ and $\cos \theta$. Recall that $\sin \left(\frac{\theta}{2}\right) = \pm \sqrt{\frac{1 - \cos \theta}{2}}$ and $\cos \left(\frac{\theta}{2}\right) = \pm \sqrt{\frac{1 + \cos \theta}{2}}$. Thus,

$$\tan \left(\frac{\theta}{2}\right) = \frac{\pm \sqrt{1 - \cos \theta}}{\sqrt{1 + \cos \theta}}$$

and so

$$u^2 = \tan^2 \left(\frac{\theta}{2}\right) = \frac{1 - \cos \theta}{1 + \cos \theta}$$

and solving this for $\cos \theta$ and $\sin \theta$ yields

$$\cos \theta = \frac{1 - u^2}{1 + u^2}, \quad \sin \theta = \pm \frac{2u}{1 + u^2}.$$

It follows that in terms of u the integral becomes

$$\int \frac{p\left(\frac{1-u^2}{1+u^2}, \pm \frac{2u}{1+u^2}\right)}{q\left(\frac{1-u^2}{1+u^2}, \pm \frac{2u}{1+u^2}\right)} \frac{2du}{1+u^2}$$

which is a rational function of u and so in theory, you might be able to find the integral from the method of partial fractions. As usual, there are no magic bullets. Even the best techniques can fail if for no other reason than our inability to factor polynomials. In calculus, there are big theorems like the fundamental theorem of calculus which have universal application and then there are many gimmicks which sometimes work. This chapter has been devoted to these gimmicks.

8.9 Using MATLAB

You can use computer algebra systems to find antiderivatives and save a lot of pain. I will explain for MATLAB and note that you can do it for any of the standard computer algebra systems. Say you want to find the obnoxious antiderivatives

$$\int \sqrt{1+3x^2} dx$$

Here is what you do in MATLAB. You must have the symbolic math package installed to do this. Remember to get to a new line, you press shift enter. You have to first write syms x to tell it that x is a variable. Enter the following. Then you press enter and it gives the answer below.

```
>>syms x
int(sqrt(1+3*x^2),x)
ans =
(3^(1/2)*asinh(3^(1/2)*x))/6 + (3^(1/2)*x*(x^2 + 1/3)^(1/2))/2
```

Note how it gives you the answer in terms of the inverse of the hyperbolic sinh. This is the meaning of the asinh. If you work it by hand, you will likely get something which looks different. You can do them all this way. Remember to write $5*6$ to indicate 5×6 . The little ^ means to write as an exponent. On my keyboard, it is above the 6. Also note that syms x makes x a variable, not a list of values.

In the distant past when I was young, we used integral tables to help us find anti-derivatives but now, you can use computer algebra. You might want to do so on some of the technical problems in the following list.

Another very easy to use algebra system is Scientific Notebook. In this, you simply type in the integral in math mode and press evaluate. I will do this now with the above integral.

$$\int \sqrt{1+3x^2} dx = \frac{1}{6} \sqrt{3} \ln \left(\sqrt{3x^2+1} + \sqrt{3}x \right) + \frac{1}{2} x \sqrt{3x^2+1}$$

Then you add the arbitrary constant. Note how it looks different. It isn't. (Why?)

8.10 Exercises

1. Give a condition on a, b , and c such that $ax^2 + bx + c$ cannot be factored as a product of two polynomials which have real coefficients.
2. Find the partial fractions expansion of the following rational functions.

(a) $\frac{2x+7}{(x+1)^2(x+2)}$

(c) $\frac{5x+1}{(x^2+1)^2(2x+3)}$

(b) $\frac{5x+1}{(x^2+1)(2x+3)}$

(d) $\frac{5x^4+10x^2+3+4x^3+6x}{(x+1)(x^2+1)^2}$

3. Find the antiderivatives

(a) $\int \frac{x^5+4x^4+5x^3+2x^2+2x+7}{(x+1)^2(x+2)} dx$

(b) $\int \frac{5x+1}{(x^2+1)(2x+3)} dx$

(c) $\int \frac{5x+1}{(x^2+1)^2(2x+3)} dx$

4. Each of $\cot \theta$, $\tan \theta$, $\sec \theta$, and $\csc \theta$ is a rational function of $\cos \theta$ and $\sin \theta$. Use the technique of substituting $u = \tan\left(\frac{\theta}{2}\right)$ to find antiderivatives for each of these.
5. Find $\int \frac{\sin \theta}{1+\sin \theta} d\theta$. **Hint:** Use the above procedure of letting $u = \tan\left(\frac{\theta}{2}\right)$ and then multiply both the top and the bottom by $(1 - \sin \theta)$ to see another way of doing it.
6. Find $\int \frac{\cos \theta + 1}{\cos \theta + 2} d\theta$ using the substitution $u = \tan\left(\frac{\theta}{2}\right)$.
7. In finding $\int \sec(x) dx$, try the substitution $u = \sin(x)$.
8. In finding $\int \csc(x) dx$ try the substitution $u = \cos(x)$.
9. Solve the following initial value problem from ordinary differential equations which is to find a function y such that

$$y'(x) = \frac{x^4 + 2x^3 + 4x^2 + 3x + 2}{x^3 + x^2 + x + 1}, y(0) = 2.$$

10. Find the antiderivatives.

(a) $\int \frac{17x-3}{(6x+1)(x-1)} dx$

(b) $\int \frac{50x^4-95x^3-20x^2-3x+7}{(5x+3)(x-2)(2x-1)} dx$ **Hint:** Notice the degree of the numerator is larger than the degree of the denominator.

(c) $\int \frac{8x^2+x-5}{(3x+1)(x-1)(2x-1)} dx$

(d) $\int \frac{3x+2}{(5x+3)(x+1)} dx$

11. Find the antiderivatives

(a) $\int \frac{52x^2+68x+46+15x^3}{(x+1)^2(5x^2+10x+8)} dx$

(b) $\int \frac{9x^2-42x+38}{(3x+2)(3x^2-12x+14)} dx$

(c) $\int \frac{9x^2-6x+19}{(3x+1)(3x^2-6x+5)} dx$

12. Solve the initial value problem $y' = f(x)$, $y(0) = 1$ for $f(x)$ equal to each of the integrands in Problem 11.

13. *Find the antiderivatives. You will need to complete the square and then make a trig. substitution.

(a) $\int \frac{1}{(3x^2+12x+13)^2} dx =$

(b) $\int \frac{1}{(5x^2+10x+7)^2} dx =$

(c) $\int \frac{1}{(5x^2-20x+23)^2} dx =$

14. Solve the initial value problem $y' = f(x)$, $y(0) = 1$ for $f(x)$ equal to each of the integrands in Problem 13.

15. Use MATLAB or some other computer algebra system to find the following antiderivatives. Some of these you really don't want to do by hand.

(a) $\int \frac{1}{1+3x^2} dx$

(b) $\int \frac{1}{\sqrt{1+5x^2}} dx$

(c) $\int \frac{x^2+2}{(x^2+2x+1)(x^2+1)} dx$

(d) $\int \sqrt{6-3x^2} dx$

(e) $\int x^7 e^{x^2} dx$

(f) $\int \sin^8(x) dx$

(g) $\int \sin^4(x) \cos^7(x) dx$

(h) $\int e^x \sin(3x) \cos(5x) dx$

(i) $\int \frac{x^2+7}{2x^5-9x^4+7x^3+14x^2-12x-8} dx$

(j) $\int \frac{1}{1+x^4} dx$

(k) $\int \frac{x^2+3x}{(x^2+x+1)^2(x^2+3)} dx$

(l) $\int \frac{x^5+2x}{(x^2+1)^2} dx$

(m) $\int \frac{1+x^2}{1+3x^4} dx$ Factor the bottom into the product of two irreducible quadratics to get the partial fractions expansion. Then stop. The remaining details are grievous.

16. Suppose $x_0 \in (a, b)$ and that f is a function which has $n + 1$ continuous derivatives on this interval. Consider the following.

$$\begin{aligned} f(x) &= f(x_0) + \int_{x_0}^x f'(t) dt = f(x_0) + (t-x) f'(t) \Big|_{x_0}^x + \int_{x_0}^x (x-t) f''(t) dt \\ &= f(x_0) + f'(x_0)(x-x_0) + \int_{x_0}^x (x-t) f''(t) dt. \end{aligned}$$

Explain the above steps and continue the process to eventually obtain Taylor's formula,

$$f(x) = f(x_0) + \sum_{k=1}^n \frac{f^{(k)}(x_0)}{k!} (x-x_0)^k + \frac{1}{n!} \int_{x_0}^x (x-t)^n f^{(n+1)}(t) dt$$

where $n! \equiv n(n-1) \cdots 3 \cdot 2 \cdot 1$ if $n \geq 1$ and $0! \equiv 1$.

17. In the above Taylor's formula, use the mean value theorem for integrals to obtain the existence of some z between x_0 and x such that

$$f(x) = f(x_0) + \sum_{k=1}^n \frac{f^{(k)}(x_0)}{k!} (x-x_0)^k + \frac{f^{(n+1)}(z)}{(n+1)!} (x-x_0)^{n+1}.$$

Hint: You might consider two cases, the case when $x > x_0$ and the case when $x < x_0$.

8.11 Videos

[volumes](#)

Chapter 9

A Few Standard Applications

As pointed out earlier, one can find the position of an object $r(t)$ by considering where it starts r_0 and knowing its velocity. Thus if the velocity $v(t)$ is known, one needs to solve the initial value problem

$$r'(t) = v(t), r(0) = r_0$$

There are many other simple problems which can be formulated as initial value problems. This chapter considers some of the standard ones. When we write dx or dy , this is rather fuzzy but it indicates a very small change in x or y . Formally,

$$\frac{dy}{dx} = f'(x), dy = f'(x) dx$$

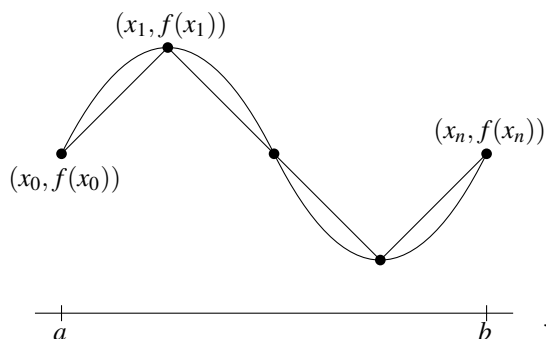
This was the way Leibniz thought of things and it is his notation used here. This can be made more rigorous, by featuring Riemann sums. However, the approach usually used in what follows is closer to what was done before Riemann. These kinds of problems and many other more complicated ones were well understood in the 1700's long before Riemann sums were introduced to make the integral more rigorous.

You might want to formulate these things in terms of approximate problems involving Riemann sums or set it up as an initial value problem. I have done this in the case of lengths, but there is a general principle here. When you do physical modeling, you don't try to achieve ultimate mathematical rigor. You make approximations and use geometrical reasoning and intuition to obtain something which can be dealt with using rigorous mathematics. Thus the topics in this chapter other than arc length are not really part of the essential mathematical content of calculus but are applications. These and other physical problems are important because they motivated the development of calculus in the first place. Mathematics is a kind of language and it is used to study various problems from other disciplines. It does not equate to these problems. It has been my experience that this distinction is often ignored or not understood, especially by people in university administration.

9.1 Lengths of Curves and Areas of Surfaces of Revolution

9.1.1 Lengths

Consider a partition of $[a, b]$, $a = x_0 < x_1 < \cdots < x_n = b$ and observe that the length of the curve between the two points $(x_i, f(x_i))$ and $(x_{i-1}, f(x_{i-1}))$ is approximately the length of the line joining these two points and that the length of the curve would be close to the sum of these as suggested in the following picture.



Thus the length of the curve would be approximately the sum of the lengths of the little straight lines in the above picture and this equals

$$\sum_{i=1}^n \sqrt{(f(x_i) - f(x_{i-1}))^2 + (x_i - x_{i-1})^2}$$

which is equal to

$$\sum_{i=1}^n \sqrt{(f'(z_i)(x_i - x_{i-1}))^2 + (x_i - x_{i-1})^2}$$

by the mean value theorem. Then this reduces to

$$\sum_{i=1}^n \sqrt{f'(z_i)^2 + 1} (x_i - x_{i-1})$$

which is a Riemann sum for the integral $\int_a^b \sqrt{1 + f'(x)^2} dx$. One would imagine that this approximation should have as the limit that which should be defined as the length of the curve.

This definition gives the right answer for the length of a straight line. To see this, consider a straight line through the points (a, b) and (c, d) where $a < c$. Then the right answer is given by the Pythagorean theorem or distance formula and is $\sqrt{(a - c)^2 + (d - b)^2}$. What is obtained from the above initial value problem? The equation of the line is $f(x) = b + \left(\frac{d-b}{c-a}\right)(x - a)$ and so $f'(x) = \left(\frac{d-b}{c-a}\right)$. Therefore, by the new procedure, the length is

$$\int_a^c \sqrt{1 + \left(\frac{d-b}{c-a}\right)^2} dx = (c - a) \sqrt{1 + \left(\frac{d-b}{c-a}\right)^2} = \sqrt{(a - c)^2 + (d - b)^2}$$

as hoped. Thus the new procedure gives the right answer in the familiar cases but it also can be used to find lengths for more general curves than straight lines. Summarizing,

Procedure 9.1.1 To find the length of the graph of the function $y = f(x)$ for $x \in [a, b]$, compute

$$\int_a^b \sqrt{1 + f'(x)^2} dx.$$

Here is another familiar example.

Example 9.1.2 Find the length of the part of the circle having radius r which is between the points $(0, r)$ and $(\frac{\sqrt{2}}{2}r, \frac{\sqrt{2}}{2}r)$.

Here the function is $f(x) = \sqrt{r^2 - x^2}$ and so $f'(x) = -x/\sqrt{r^2 - x^2}$. Therefore, the length is

$$\int_0^{\pi/4} \sqrt{1 + \left(-x/\sqrt{r^2 - x^2}\right)^2} dx = \int_0^{\pi/4} r \sqrt{\left(\frac{1}{r^2 - x^2}\right)} dx$$

Using a trig substitution $x = r \sin \theta$, it follows $dx = r \cos(\theta) d\theta$ and so

$$\int \frac{r}{\sqrt{r^2 - x^2}} dx = \int \frac{1}{\sqrt{1 - \sin^2 \theta}} r \cos(\theta) d\theta = r \int d\theta = r\theta + C$$

Hence changing back to the variable x it follows an antiderivative is

$$l(x) = r \arcsin\left(\frac{x}{r}\right)$$

Then the length is

$$r \arcsin\left(\frac{r}{r}\right) - r \arcsin\left(\frac{1}{r} \frac{\sqrt{2}}{2} r\right) = r \frac{\pi}{2} - r \frac{\pi}{4} = r \frac{\pi}{4}.$$

Note this gives the length of one eighth of the circle and so from this the length of the whole circle should be $2r\pi$. Here is another example

Example 9.1.3 Find the length of the graph of $y = x^2$ between $x = 0$ and $x = 1$.

Here $f'(x) = 2x$ and so the initial value problem to be solved is

$$\frac{dl}{dx} = \sqrt{1 + 4x^2}, \quad l(0) = 0.$$

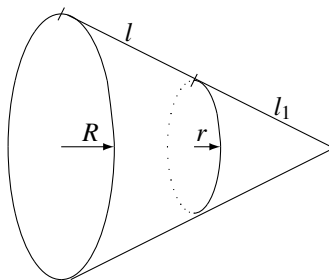
Thus, in terms of the definite integral, the length of this curve is

$$\int_0^1 \sqrt{1 + 4x^2} dx = \frac{1}{2} \sqrt{5} - \frac{1}{4} \ln(-2 + \sqrt{5}) = \frac{1}{2} \sqrt{5} + \frac{1}{4} \ln(\sqrt{5} + 2)$$

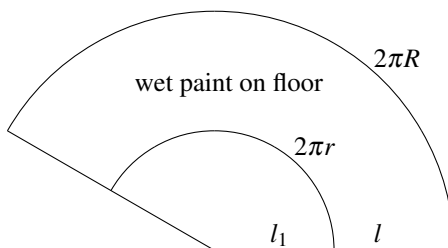
To find an antiderivative, you use the trig. substitution $2x = \tan u$ so $dx = \frac{1}{2} (\sec^2 u) du$. Then you find the antiderivative in terms of u and change back to x by using an appropriate triangle as described earlier.

9.1.2 Surfaces of Revolution

The problem of finding the surface area of a solid of revolution is closely related to that of finding the length of a graph. First consider the following picture of the frustum of a cone in which it is desired to find the lateral surface area. In this picture, the frustum of the cone is the left part which has an l next to it and the lateral surface area is this part of the area of the cone.



To do this, imagine painting the sides and rolling the shape on the floor for exactly one revolution. The wet paint would make the following shape.



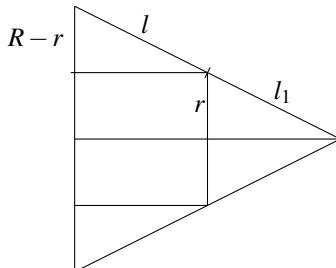
What would be the area of this wet paint? Its area would be the difference between the areas of the two sectors shown, one having radius l_1 and the other having radius $l + l_1$. Both of these have the same central angle equal to

$$\frac{2\pi R}{2\pi(l + l_1)} 2\pi = \frac{2\pi R}{l + l_1}.$$

Therefore, Theorem 2.3.12, this area is

$$(l + l_1)^2 \frac{\pi R}{(l + l_1)} - l_1^2 \frac{\pi R}{(l + l_1)} = \pi R l \frac{l + 2l_1}{l + l_1}$$

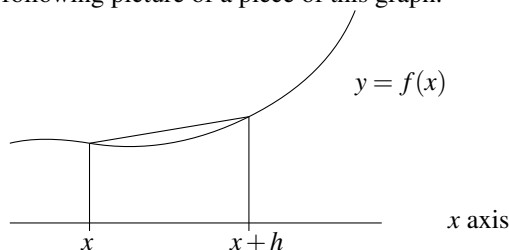
The view from the side is



and so by similar triangles, $l_1 = lr/(R-r)$. Therefore, substituting this into the above, the area of this frustum is

$$\pi Rl \frac{l + 2\left(\frac{lr}{R-r}\right)}{l + \left(\frac{lr}{R-r}\right)} = \pi l(R+r) = 2\pi l \left(\frac{R+r}{2}\right).$$

Now consider a function f , defined on an interval $[a, b]$ and suppose it is desired to find the area of the surface which results when the graph of this function is revolved about the x axis. Consider the following picture of a piece of this graph.



Let $A(x)$ denote the area which results from revolving the graph of the function restricted to $[a, x]$ about the x axis. Then from the above formula for the area of a frustum,

$$\frac{A(x+h) - A(x)}{h} \approx 2\pi \frac{1}{h} \sqrt{h^2 + (f(x+h) - f(x))^2} \left(\frac{f(x+h) + f(x)}{2} \right)$$

where \approx denotes that these are close to being equal and the approximation gets increasingly good as $h \rightarrow 0$. Therefore, rewriting this a little yields

$$\frac{A(x+h) - A(x)}{h} \approx 2\pi \sqrt{1 + \left(\frac{f(x+h) - f(x)}{h} \right)^2} \left(\frac{f(x+h) + f(x)}{2} \right)$$

Therefore, taking the limit as $h \rightarrow 0$, and using $A(a) = 0$, this yields the following initial value problem for A which can be used to find the area of a surface of revolution.

$$A'(x) = 2\pi f(x) \sqrt{1 + f'(x)^2}, \quad A(a) = 0.$$

What would happen if you revolved about the y axis? I will leave it to you to verify this would lead to the initial value problem

$$A'(x) = 2\pi x \sqrt{1 + f'(x)^2}, \quad A(a) = 0.$$

As before, this results in the following simple procedure for finding the surface area of a surface of revolution.

Procedure 9.1.4 To find the surface area of a surface obtained by revolving the graph of $y = f(x)$ for $x \in [a, b]$ about the x axis, compute

$$\int_a^b 2\pi f(x) \sqrt{1 + f'(x)^2} dx$$

Similarly, to get the area of the graph rotated about the y axis, compute

$$\int_a^b 2\pi x \sqrt{1 + f'(x)^2} dx.$$

Example 9.1.5 Find the surface area of the surface obtained by revolving the function $y = r$ for $x \in [a, b]$ about the x axis. Of course this is just the cylinder of radius r and height $b - a$ so this area should equal $2\pi r(b - a)$. (Imagine painting it and rolling it on the floor and then taking the area of the rectangle which results.)

Using the above initial value problem, solve

$$A'(x) = 2\pi r \sqrt{1 + 0^2}, \quad A(a) = 0.$$

The solution is $A(x) = 2\pi r(x - a)$. Therefore, $A(b) = 2\pi r(b - a)$ as expected.

Example 9.1.6 Find the surface area of a sphere of radius r .

Here the function involved is $f(x) = \sqrt{r^2 - x^2}$ for $x \in [-r, r]$ and it is to be revolved about the x axis. In this case

$$f'(x) = \frac{-x}{\sqrt{r^2 - x^2}}$$

and so, by the procedure described above, the surface area is

$$\int_{-r}^r 2\pi \sqrt{r^2 - x^2} \sqrt{1 + \frac{x^2}{r^2 - x^2}} dx = 4r^2\pi$$

9.2 Exercises

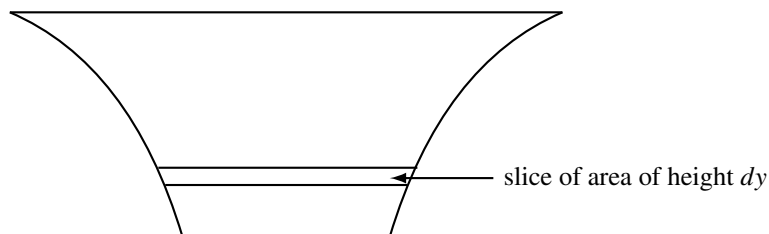
- Find the length of the graph of $y = \ln(\cos x)$ for $x \in [0, \pi/4]$.
- The curve defined by $y = \ln(\cos x)$ for $x \in [0, 1]$ is revolved about the y axis. Find an integral for the area of the surface of revolution.
- Find the length of the graph of $y = x^{1/2} - \frac{x^{3/2}}{3}$ for $x \in [1, 3]$.
- The graph of the function $y = x^3$ is revolved about the x axis for $x \in [0, 1]$. Find the area of the resulting surface of revolution.
- The graph of the function $y = x^3$ is revolved about the y axis for $x \in [0, 1]$. Find the area of the resulting surface of revolution. **Hint:** Formulate this in terms of x and use a change of variables.
- The graph of the function $y = \ln x$ is revolved about the y axis for $x \in [1, 2]$. Find the area of the resulting surface of revolution. **Hint:** Consider x as a function of y .
- The graph of the function $y = \ln x$ is revolved about the x axis for $x \in [1, 2]$. Find the area of the resulting surface of revolution. If you cannot do the integral, set it up.
- Find the length of $y = \cosh(x)$ for $x \in [0, 1]$.
- Find the length of $y = 2x^2 - \frac{1}{16} \ln x$ for $x \in [1, 2]$.
- The curve defined by $y = 2x^2 - \frac{1}{16} \ln x$ for $x \in [1, 2]$ is revolved about the y axis. Find the area of the resulting surface of revolution.

11. Find the length of $y = x^2 - \frac{1}{8} \ln x$ for $x \in [1, 2]$.
12. The curve defined by $y = x^2 - \frac{1}{8} \ln x$ for $x \in [1, 2]$ is revolved about the y axis. Find the area of the resulting surface of revolution.
13. The curve defined by $y = \cosh(x)$ for $x \in [0, 1]$ is revolved about the x axis. Find the area of the resulting surface of revolution.
14. The curve defined by $y = \cosh(x)$ for $x \in [0, 1]$ is revolved about the line $y = -3$. Find the area of the resulting surface of revolution.
15. For a a positive real number, find the length of $y = \frac{ax^2}{2} - \frac{1}{4a} \ln x$ for $x \in [1, 2]$. Of course your answer should depend on a .
16. The graph of the function $y = x^2$ for $x \in [0, 1]$ is revolved about the x axis. Find the area of the surface of revolution.
17. The graph of the function $y = \sqrt{x}$ for $x \in [0, 1]$ is revolved about the y axis. Find the area of the surface of revolution. **Hint:** Switch x and y and then use the previous problem.
18. The graph of the function $y = x^{1/2} - \frac{x^{3/2}}{3}$ is revolved about the y axis. Find the area of the surface of revolution if $x \in [0, 2]$.
19. The graph of the function $y = \sinh x$ for $x \in [0, 1]$ is revolved about the x axis. Find the area of the surface of revolution.
20. * The ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ is revolved about the x axis. Find the area of the surface of revolution.
21. Find the length of the graph of $y = \frac{2}{3}(x-1)^{3/2}$ for $x \in [2, 3]$.
22. The curve defined by $y = \frac{2}{3}(x-1)^{3/2}$ for $x \in [1, 2]$ is revolved about the y axis. Find the area of the resulting surface of revolution.
23. Suppose $f'(x) = \sqrt{\sec^2 x - 1}$ and $f(0) = 0$. Find the length of the graph of $y = f(x)$ for $x \in [0, 1]$.
24. The curve defined by $y = f(x)$ for $x \in [0, \pi]$ is revolved about the y axis where $f'(x) = \sqrt{(2 + \sin x)^2 - 1}$, $f(0) = 1$. Find the area of the resulting surface of revolution.
25. Revolve $y = 1/x$ for $x \in [1, R]$ about the x axis. Find the area of this surface of revolution. Now show the limit of what you got as $R \rightarrow \infty$ does not exist. Next find the volume of this solid of revolution. Show the limit of this as $R \rightarrow \infty$ is finite. This infinite solid has infinite area but finite volume.
26. The surface area of a sphere of radius r was shown to be $4\pi r^2$. Note that if $V(r) = \frac{4}{3}\pi r^3$, then $V'(r)$ equals the area of the sphere. Why is this reasonable based on geometrical considerations?

9.3 Force on a Dam and Work

9.3.1 Force on a Dam

Imagine you are a fish swimming in a lake behind a dam and you are interested in the total force acting on the dam. The following picture is what you would see.



The reason you would be interested in that long thin slice of area having essentially the same depth, say at y feet is because the pressure in the water at that depth is constant and equals $62.5y$ pounds per square foot¹. Therefore, the total force the water exerts on the long thin slice is

$$dF = 62.5yL(y)dy$$

where $L(y)$ denotes the length of the slice. Therefore, the total force on the dam up to depth y is obtained as a solution to the initial value problem

$$\frac{dF}{dy} = 62.5yL(y), \quad F(0) = 0.$$

Example 9.3.1 Suppose the width of a dam at depth y feet equals $L(y) = 1000 - y$ and its depth is 500 feet. Find the total force in pounds exerted on the dam.

From the above, this is obtained as the solution to the initial value problem

$$\frac{dF}{dy} = 62.5y(1000 - y), \quad F(0) = 0$$

which is $F(y) = -20.83y^3 + 31250y^2$. The total force on the dam would be

$$F(500) = -20.83(500)^3 + 31250(500)^2 = 5,208,750,000.0$$

pounds. In tons this is 2,604,375. That is a lot of force.

9.3.2 Work

Now suppose you are pumping water from a tank of depth d to a height of H feet above the top of the water in the tank. Suppose also that at depth y below the surface, the area of a cross section having constant depth is $A(y)$. The total weight of a slice of water having thickness dy at this depth is $62.5A(y)dy$ and the pump needs to lift this weight a distance

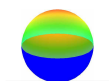
¹That this is so comes from an assumption that water is incompressible and the use of topics in multivariable calculus. Here we will simply use this fact, but it is derived later in the book. It was first observed experimentally by Blaise Pascal, a very important French theologian, philosopher, physicist and mathematician of the 1600's.

of $y + H$ feet. Therefore, the work done is $dW = (y + H) 62.5A(y) dy$. An initial value problem for the work done to pump the water down to a depth of y feet would be

$$\frac{dW}{dy} = (y + H) 62.5A(y), \quad W(0) = 0.$$

The reason for the initial condition is that the pump has done no work to pump no water. If the weight of the fluid per cubic foot were different than 62.5 you would do the same things but replace the number.

Example 9.3.2 A spherical storage tank sitting on the ground having radius 30 feet is half filled with a fluid which weighs 50 pounds per cubic foot. How much work is done to pump this fluid to a height of 100 feet?



Letting r denote the radius of a cross section y feet below the level of the fluid, $r^2 + y^2 = 900$. Therefore, $r = \sqrt{900 - y^2}$. It follows the area of the cross section at depth y is $\pi(900 - y^2)$. Here $H = 70$ and so the initial value problem to solve is

$$\frac{dW}{dy} = (y + 70) 50\pi(900 - y^2), \quad W(0) = 0.$$

Therefore, $W(y) = 50\pi\left(-\frac{1}{4}y^4 - \frac{70}{3}y^3 + 450y^2 + 63000y\right)$ and the total work in foot pounds equals

$$W(30) = 50\pi\left(-\frac{1}{4}(30)^4 - \frac{70}{3}(30)^3 + 450(30)^2 + 63000(30)\right) = 73,125,000\pi$$

In general, the work done by a constant force in a straight line equals the product of the force times the distance over which it acts. If the force is varying with respect to position, then you have to use calculus to compute the work. For now, consider the following examples.

Example 9.3.3 A 500 pound safe is lifted 10 feet. How much work is done?

The work is $500 \times 10 = 5000$ foot pounds.

Example 9.3.4 The force needed to stretch a spring x feet past its equilibrium position is kx . This is known as Hooke's law and is a good approximation as long as the spring is not stretched too far. If $k = 3$, how much work is needed to stretch the spring a distance of 2 feet beyond its equilibrium position? The constant k is called the spring constant. Different springs would have different spring constants. The units on k are pounds/foot.

This is a case of a variable force. To stretch the spring from x to $x + dx$ requires $3xdx$ foot pounds of work. Therefore, letting W denote the work up till time x , $dW = 3xdx$ and so the initial value problem is

$$\frac{dW}{dx} = 3x, \quad W(0) = 0.$$

Thus $W(2) = \frac{3}{2}(2^2) = 6$ foot pounds because an antiderivative for $3x$ is $\frac{3}{2}x^2$. In terms of the definite integral, this is written as $\int_0^2 3xdx$.

9.4 Using MATLAB

Sometimes when you do applications, you end up with an integral you can't evaluate because you don't know how to find an antiderivative. When this occurs, you need to use a numerical method. This is a long story best left to numerical analysis courses, but you can easily get the answer numerically with computer algebra. Suppose you want to find the really obnoxious integral $\int_0^5 \sin(x) \exp(-x^2) dx$. In MATLAB, you would enter the following.

```
f=@(x)sin(x).*exp(-x.^2);
```

```
integral(f,0,5)
```

Then you press enter and it gives.

```
ans =
```

```
0.4244
```

The first line defines the function and the second tells it to find the integral mentioned above. You have to use `.*` because you are dealing with lists of numbers and you want to do the multiplication to corresponding entries. MATLAB is like that. It will see `x` as a list of numbers and `sin(x)` as a list of numbers obtained from taking the sine of each number in the list for `x`. It is similar for `exp`.

There is a discussion of numerical integration schemes in Problem 25 on Page 198. However, a rudimentary integration scheme is the Riemann sum. The versions in the above problem are much better and what MATLAB uses is still more sophisticated.

If you have Scientific Notebook, it is even easier. You simply type

$$\int_0^5 \sin(x) \exp(-x^2) dx$$

and then press evaluate numerically $\stackrel{#}{=}$? on the toolbar. The result is

$$\int_0^5 \sin(x) \exp(-x^2) dx = 0.42444$$

Actually, this software is built on Mupad which is a part of the symbolic math package of MATLAB.

9.5 Exercises

1. The main span of the Portage Lake lift bridge² weighs 4,400,000 pounds. How much work is done in raising this main span to a height of 100 feet?
2. A cylindrical storage tank having radius 20 feet and length 40 feet is filled with a fluid which weighs 50 pounds per cubic foot. This tank is lying on its side on the ground. Find the total force acting on the ends of the tank by the fluid.

²This is the heaviest lift bridge in the world. It joins the towns of Houghton and Hancock in the upper peninsula of Michigan spanning Portage lake. It provides 250 feet of clear channel for ships and can provide as much as 100 feet of vertical clearance. The lifting machinery is at the top of two massive towers 180 feet above the water. Aided by 1,100 ton counter weights on each tower, sixteen foot gears pull on 42 cables to raise the bridge. This usually creates impressive traffic jams on either side of the lake. The motion up and down of this span is quite slow.

3. Suppose the tank in Problem 2 is filled to a depth of 8 feet. Find an integral for the work needed to pump the fluid to a height of 50 feet.
4. A conical hole is filled with water which has weight 62.5 pounds per cubic foot. If the depth of the hole is 20 feet and the radius of the hole is 10 feet, how much work is needed to pump the water to a height of 10 feet above the ground?
5. Suppose the spring constant is 2 pounds per foot. Find the work needed to stretch the spring 3 feet beyond equilibrium.
6. A 20 foot chain lies on the ground. It weighs 5 pounds per foot. How much work is done to lift one end of the chain to a height of 20 feet?
7. A 200 foot chain dangles from the top of a tall building. How much work is needed to haul it to the top of the building if it weighs 1 pound per foot?
8. A dam 500 feet high has a width at depth y equal to $4000 - 2y$ feet. What is the total force on the dam if it is filled?
9. *When the bucket is filled with water it weighs 30 pounds and when empty it weighs 2 pounds and the person on top of a 100 foot building exerts a constant force of 40 pounds. The bucket is full at the bottom but leaks at the rate of .1 cubic feet per second. How much work does the person on the top of the building do in lifting the bucket to the top? Will the bucket be empty when it reaches the top? You can use Newton's law that force equals mass times acceleration. You can neglect the weight of the rope.
10. In the situation of the above problem, suppose the person on the top maintains a constant velocity of 1 foot per second and the bucket leaks at the rate of .1 pound per second. How much work does he do and is the bucket empty when it reaches the top?
11. A silo is 10 feet in diameter and at a height of 30 feet there is a hemispherical top. The silage weighs 10 pounds per cubic foot. How much work was done in filling it to the very top?
12. A cylindrical storage tank having radius 10 feet is filled with water to a depth of 20 feet. If the storage tank stands upright on its circular base, what is the total force the water exerts on the sides of the tank? **Hint:** The pressure in the water at depth y is $62.5y$ pounds per square foot.
13. A spherical storage tank having radius 10 feet is filled with water. What is the total force the water exerts on the storage tank? **Hint:** The pressure in the water at depth y is $62.5y$ consider the area corresponding to a slice at height y . This is a surface of revolution and you know how to deal with these. The area of this slice times the pressure gives the total force acting on it.
14. A water barrel which is 11 inches in radius and 34 inches high is filled with water. If it is standing on end, what is the total force acting on the circular sides of the barrel?
15. Find the total force acting on the circular sides of the cylinder in Problem 2.

16. A cylindrical tank having radius 10 feet is contains water which weight 62.5 pounds per cubic foot. Find the force on one end of this tank if it is filled to a depth of y feet.
17. Here is a calculator problem. In the above problem, to what depth may the tank be filled if the total force on an end is not to exceed 40000 pounds?
18. The force on a satellite of mass m slugs in pounds is $\frac{mk}{r^2}$ where k is approximately $k = 1.42737408 \times 10^{16}$ and r is the distance from the center of the earth. Assuming the radius of the earth is 4000 miles, find the work in foot pounds needed to place a satellite weighing 500 pounds on the surface of the earth into an orbit 18,000 miles above the surface of the earth. You should use a calculator on this problem.
19. A **regular Sturm Liouville problem** involves the differential equation, for an unknown function of x which is denoted here by y ,

$$(p(x)y')' + (\lambda q(x) + r(x))y = 0, \quad x \in [a, b]$$

and it is assumed that $p(t), q(t) > 0$ for any t along with boundary conditions,

$$\begin{aligned} C_1 y(a) + C_2 y'(a) &= 0, \\ C_3 y(b) + C_4 y'(b) &= 0 \end{aligned}$$

where

$$C_1^2 + C_2^2 > 0, \text{ and } C_3^2 + C_4^2 > 0.$$

There is an immense theory connected to these important problems. The constant, λ is called an eigenvalue. Show that if y is a solution to the above problem corresponding to $\lambda = \lambda_1$ and if z is a solution corresponding to $\lambda = \lambda_2 \neq \lambda_1$, then

$$\int_a^b q(x) y(x) z(x) dx = 0. \quad (9.1)$$

Hint: Do something like this:

$$\begin{aligned} (p(x)y')' z + (\lambda_1 q(x) + r(x)) yz &= 0, \\ (p(x)z')' y + (\lambda_2 q(x) + r(x)) zy &= 0. \end{aligned}$$

Now subtract and either use integration by parts or show

$$(p(x)y')' z - (p(x)z')' y = ((p(x)y') z - (p(x)z') y)')$$

and then integrate. Use the boundary conditions to show that $y'(a)z(a) - z'(a)y(a) = 0$ and $y'(b)z(b) - z'(b)y(b) = 0$. The formula, 9.1 is called an orthogonality relation and it makes possible an expansion in terms of certain functions called eigenfunctions.

Chapter 10

Improper Integrals and Stirling's Formula

10.1 Stirling's Formula

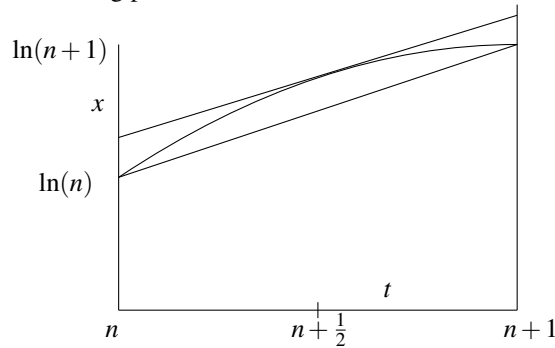
In this section is an elementary approach to Stirlings formula. This formula is an asymptotic approximation for $n!$. It is quite old dating to about 1730. The approach followed here is like the one in the Calculus book of Courant found in the references. Later I will give a different one found in [26]. See also [8].

To begin with is a simple lemma which really depends on the shape of the graph of $t \rightarrow \ln t$.

Lemma 10.1.1 *For n a positive integer,*

$$\frac{1}{2} (\ln(n+1) + \ln(n)) \leq \int_n^{n+1} \ln(t) dt \leq \ln\left(n + \frac{1}{2}\right) \quad (10.1)$$

Proof: Consider the following picture.



There are two trapezoids, the area of the larger one is larger than $\int_n^{n+1} \ln(t) dt$ and the area of the smaller being smaller than this integral. The equation of the line which forms the top of the large trapezoid is

$$y - \ln\left(n + \frac{1}{2}\right) = \frac{1}{n + \frac{1}{2}} \left(x - \left(n + \frac{1}{2}\right)\right)$$

Thus the area of the large trapezoid is obtained by averaging the two vertical sides and multiplying by the length of the base which is 1. This is easily found to be $\ln\left(n + \frac{1}{2}\right)$. Then the area of the smaller trapezoid is obtained also as the average of the two vertical sides times the length of the base which is $\frac{1}{2}(\ln(n+1) + \ln(n))$. ■

Thus a lower approximation for $\int_1^n \ln(t) dt$, denoted as T_n is

$$T_n \equiv \sum_{k=1}^{n-1} \frac{1}{2} (\ln(k) + \ln(k+1))$$

Then, from the above lemma,

$$\begin{aligned} \int_1^n \ln(t) dt - T_n &= \sum_{k=1}^{n-1} \int_k^{k+1} \ln(t) dt - \sum_{k=1}^{n-1} \frac{1}{2} (\ln(k) + \ln(k+1)) \\ &\leq \sum_{k=1}^{n-1} \ln\left(k + \frac{1}{2}\right) - \sum_{k=1}^{n-1} \frac{1}{2} (\ln(k) + \ln(k+1)) \\ &= \sum_{k=1}^{n-1} \frac{1}{2} \left(\ln\left(k + \frac{1}{2}\right) - \ln(k) \right) - \sum_{k=1}^{n-1} \frac{1}{2} \left(\ln(k+1) - \ln\left(k + \frac{1}{2}\right) \right) \\ &\leq \sum_{k=1}^{n-1} \frac{1}{2} \left(\ln(k) - \ln\left(k - \frac{1}{2}\right) \right) - \sum_{k=1}^{n-1} \frac{1}{2} \left(\ln(k+1) - \ln\left(k + \frac{1}{2}\right) \right) \\ &= \sum_{k=0}^{n-2} \frac{1}{2} \left(\ln(k+1) - \ln\left(k + \frac{1}{2}\right) \right) - \sum_{k=1}^{n-1} \frac{1}{2} \left(\ln(k+1) - \ln\left(k + \frac{1}{2}\right) \right) \\ &= \frac{1}{2} \left(\ln(1) - \ln\left(\frac{1}{2}\right) \right) - \frac{1}{2} \left(\ln(n) - \ln\left(n - \frac{1}{2}\right) \right) \leq \frac{\ln(2)}{2} \end{aligned}$$

Now this shows that $\{\int_1^n \ln(t) dt - T_n\}_{n=1}^{\infty}$ is an increasing sequence bounded above and so it must converge to some real number α .

$$\begin{aligned} \exp(T_n) &= \prod_{k=1}^{n-1} \exp\left(\frac{1}{2} (\ln(k) + \ln(k+1))\right) = \prod_{k=1}^{n-1} (k(k+1))^{1/2} \\ &= (1 \cdot 2)^{1/2} (2 \cdot 3)^{1/2} \dots ((n-1) \cdot n)^{1/2} = (n-1)! \sqrt{n} = n! n^{-1/2} \end{aligned}$$

Therefore, doing the integral $\int_1^n \ln(t) dt$ and taking the exponential of the expression,

$$\lim_{n \rightarrow \infty} \exp((n \ln(n) - n) - T_n) = \lim_{n \rightarrow \infty} \frac{e^{(n \ln(n) - n)}}{n^{-1/2} n!} = \lim_{n \rightarrow \infty} \frac{n^{n+1/2} e^{-n}}{n!} = e^{\alpha}$$

This has proved the following lemma.

Lemma 10.1.2 *There exists a positive number c such that*

$$\lim_{n \rightarrow \infty} \frac{n!}{n^{n+(1/2)} e^{-n} c} = 1.$$

In many applications, the above is enough. However, the constant can be found. There are various ways to show that this constant c equals $\sqrt{2\pi}$. The version given here also includes a formula which is interesting for its own sake.

Using integration by parts, it follows that whenever n is a positive integer larger than 1,

$$\int_0^{\pi/2} \sin^n(x) dx = \frac{n-1}{n} \int_0^{\pi/2} \sin^{n-2}(x) dx$$

Lemma 10.1.3 For $m \geq 1$,

$$\begin{aligned} \int_0^{\pi/2} \sin^{2m}(x) dx &= \frac{(2m-1) \cdots 1}{2m(2m-2) \cdots 2} \frac{\pi}{2} \\ \int_0^{\pi/2} \sin^{2m+1}(x) dx &= \frac{(2m)(2m-2) \cdots 2}{(2m+1)(2m-1) \cdots 3} \end{aligned}$$

Proof: Consider the first formula in the case where $m = 1$. From beginning calculus,

$$\int_0^{\pi/2} \sin^2(x) dx = \frac{\pi}{4} = \frac{1}{2} \frac{\pi}{2}$$

so the formula holds in this case. Suppose it holds for m . Then from the above reduction identity and induction,

$$\begin{aligned} \int_0^{\pi/2} \sin^{2m+2}(x) dx &= \frac{2m+1}{2(m+1)} \int_0^{\pi/2} \sin^{2m}(x) dx \\ &= \frac{2m+1}{2(m+1)} \frac{(2m-1) \cdots 1}{2m(2m-2) \cdots 2} \frac{\pi}{2}. \end{aligned}$$

The second claim is proved similarly. ■

Then using the reduction identity and the above,

$$\begin{aligned} \frac{2m+1}{2m} &\geq \frac{\int_0^{\pi/2} \sin^{2m}(x) dx}{\frac{2m}{2m+1} \int_0^{\pi/2} \sin^{2m-1}(x) dx} = \frac{\int_0^{\pi/2} \sin^{2m}(x) dx}{\int_0^{\pi/2} \sin^{2m+1}(x) dx} = \\ &= \frac{\pi}{2} (2m+1) \frac{(2m-1)^2 (2m-3)^2 \cdots 1}{2^{2m} (m!)^2} \geq 1 \end{aligned}$$

It follows from the squeezing theorem that

$$\lim_{m \rightarrow \infty} \frac{1}{2m+1} \frac{2^{2m} (m!)^2}{(2m-1)^2 (2m-3)^2 \cdots 1} = \frac{\pi}{2}$$

This exceedingly interesting formula is Wallis' formula.

Now multiply both the top and the bottom of the expression on the left by

$$(2m)^2 (2(m-1))^2 \cdots 2^2$$

which is $2^{2m} (m!)^2$. This is another version of the Wallis formula.

$$\frac{\pi}{2} = \lim_{m \rightarrow \infty} \frac{2^{2m}}{2m+1} \frac{2^{2m} (m!)^2 (m!)^2}{((2m)!)^2}$$

It follows that

$$\sqrt{\frac{\pi}{2}} = \lim_{m \rightarrow \infty} \frac{2^{2m}}{\sqrt{2m+1}} \frac{(m!)^2}{(2m)!} = \lim_{m \rightarrow \infty} \frac{2^{2m}}{\sqrt{2m}} \frac{(m!)^2}{(2m)!} \quad (10.2)$$

Now with this result, it is possible to find c in Stirling's formula. Recall

$$\lim_{m \rightarrow \infty} \frac{m!}{m^{m+(1/2)} e^{-m} c} = 1 = \lim_{m \rightarrow \infty} \frac{m^{m+(1/2)} e^{-m} c}{m!}$$

In particular, replacing m with $2m$,

$$\lim_{m \rightarrow \infty} \frac{(2m)!}{(2m)^{2m+(1/2)} e^{-2m} c} = \lim_{m \rightarrow \infty} \frac{(2m)^{2m+(1/2)} e^{-2m} c}{(2m)!} = 1$$

Therefore, from 10.2, $\sqrt{\frac{\pi}{2}} =$

$$\begin{aligned} \lim_{m \rightarrow \infty} \frac{2^{2m}}{\sqrt{2m}} \frac{\left(\frac{m^{m+(1/2)} e^{-m} c}{m!} \right)^2 (m!)^2}{\left(\frac{(2m)^{2m+(1/2)} e^{-2m} c}{2m!} \right) (2m)!} &= \lim_{m \rightarrow \infty} \frac{2^{2m}}{\sqrt{2m}} \frac{\left(m^{m+(1/2)} e^{-m} c \right)^2}{\left((2m)^{2m+(1/2)} e^{-2m} c \right)} \\ &= c \lim_{m \rightarrow \infty} \frac{2^{2m}}{\sqrt{2m}} \frac{m^{2m+1}}{2^{2m+1/2} (m^{2m+(1/2)})} = c \lim_{m \rightarrow \infty} \frac{1}{2} \frac{m^{2m+1}}{m^{2m+1}} = \frac{c}{2} \end{aligned}$$

so $c = \sqrt{2\pi}$. This proves Stirling's formula.

Theorem 10.1.4 *The following formula holds.*

$$\lim_{m \rightarrow \infty} \frac{m!}{m^{m+(1/2)} e^{-m}} = \sqrt{2\pi}$$

10.2 The Gamma Function

This belongs to a larger set of ideas concerning improper integrals, but the main reason for these ideas are important examples like the Gamma function or Laplace transforms. General theory is much better understood in the context of the Lebesgue integral. Therefore, the presentation is centered on these examples. The Riemann integral only is defined for bounded functions which are defined on a bounded interval. If this is not the case, then the integral has not been defined. Of course, just because the function is bounded does not mean the integral exists as mentioned above, but if it is not bounded or if it is defined on an infinite interval, then no definition has been given. However, one can consider limits of Riemann integrals. The following definition pertains to the Gamma function and Laplace transforms.

Definition 10.2.1 *We say that f defined on $[0, \infty)$ is improper Riemann integrable if it is Riemann integrable on $[\delta, R]$ for each $R > 1 > \delta > 0$ and the following limits exist.*

$$\int_0^\infty f(t) dt \equiv \lim_{\delta \rightarrow 0+} \int_\delta^1 f(t) dt + \lim_{R \rightarrow \infty} \int_1^R f(t) dt$$

The gamma function is defined by

$$\Gamma(\alpha) \equiv \int_0^\infty e^{-t} t^{\alpha-1} dt$$

whenever $\alpha > 0$.

The following lemma comes from Proposition 4.12.12 about limits of increasing or decreasing functions.

Lemma 10.2.2 *The gamma function exists because the limits in the above definition exists for each $\alpha > 0$.*

Proof: Note first that as $\delta \rightarrow 0+$, the Riemann integrals $\int_{\delta}^1 e^{-t} t^{\alpha-1} dt$ increase. Thus $\lim_{\delta \rightarrow 0+} \int_{\delta}^1 e^{-t} t^{\alpha-1} dt$ either is $+\infty$ or it will converge to the least upper bound thanks to completeness of \mathbb{R} . See Proposition 4.12.12. However, $e^{-t} t^{\alpha-1} \leq t^{\alpha-1}$ and

$$\int_{\delta}^1 t^{\alpha-1} dt = \frac{t^{\alpha}}{\alpha} \Big|_{\delta}^1 = \frac{1}{\alpha} - \frac{\delta^{\alpha}}{\alpha} \leq \frac{1}{\alpha}$$

so the limit of these integrals exists because they are bounded above. Also $e^{-t} t^{\alpha-1} \leq C e^{-(t/2)}$ for suitable C if $t > 1$. This is obvious if $\alpha - 1 < 0$ and in the other case it is also clear because

$$0 < \frac{e^{-t} t^{\alpha-1}}{e^{-(t/2)}} \leq e^{-t/2} t^m, \text{ where } m \text{ is an integer larger than } \alpha - 1$$

Now apply L'Hopital's rule to conclude that the limit of this expression is 0 as $t \rightarrow \infty$. Thus the quotient $\frac{e^{-t} t^{\alpha-1}}{e^{-(t/2)}}$ is less than some constant C .

$$\int_1^R e^{-t} t^{\alpha-1} dt \leq \int_1^R C e^{-(t/2)} dt \leq 2C e^{(-1/2)} - 2C e^{(-R/2)} \leq 2C e^{(-1/2)}$$

Thus these integrals also converge as $R \rightarrow \infty$ because they are increasing in R and bounded above. Hence they converge to $\sup \left\{ \int_1^R e^{-t} t^{\alpha-1} dt : R > 1 \right\}$. It follows that $\Gamma(\alpha)$ makes sense. ■

The argument also implies the following proposition. Absolute convergence implies convergence.

Proposition 10.2.3 *If $f \geq 0$, then $\int_a^{\infty} f(t) dt$ exists if the partial integrals $\int_a^R f(t) dt$ are bounded above independent of R . Also $\int_a^{\infty} f(t) dt$ exists if $\int_a^{\infty} |f(t)| dt$ exists.*

Proof: The first part is just like what was done with the gamma function. As to the second part, consider $f_+(t) \equiv \frac{|f(t)|+f(t)}{2}$, $f_-(t) \equiv \frac{|f(t)|-f(t)}{2}$. These are both nonnegative and if $\int_a^{\infty} |f| dt$ exists, then

$$\int_a^R f_+ dt \leq \int_a^{\infty} |f| dt, \quad \int_a^R f_- dt \leq \int_a^{\infty} |f| dt$$

and so the first part implies $\lim_{R \rightarrow \infty} \int_a^R f_+ dt$ and $\lim_{R \rightarrow \infty} \int_a^R f_- dt$ both exist. Hence

$$\int_a^R f dt = \int_a^R f_+ dt - \int_a^R f_- dt$$

also must have a limit as $R \rightarrow \infty$. ■

This gamma function has some fundamental properties described in the following proposition. In case the improper integral exists, we can obviously compute it in the form

$$\lim_{\delta \rightarrow 0+} \int_{\delta}^{1/\delta} f(t) dt$$

which is used in what follows. Thus also the usual algebraic properties of the Riemann integral are inherited by the improper integral.

Proposition 10.2.4 For n a positive integer, $n! = \Gamma(n+1)$. In general, the following identity holds. $\Gamma(1) = 1, \Gamma(\alpha+1) = \alpha\Gamma(\alpha)$

Proof: First of all, $\Gamma(1) = \lim_{\delta \rightarrow 0} \int_{\delta}^{\delta^{-1}} e^{-t} dt = \lim_{\delta \rightarrow 0} (e^{-\delta} - e^{-(\delta^{-1})}) = 1$. Next, for $\alpha > 0$,

$$\begin{aligned}\Gamma(\alpha+1) &= \lim_{\delta \rightarrow 0} \int_{\delta}^{\delta^{-1}} e^{-t} t^{\alpha} dt = \lim_{\delta \rightarrow 0} \left[-e^{-t} t^{\alpha} \Big|_{\delta}^{\delta^{-1}} + \alpha \int_{\delta}^{\delta^{-1}} e^{-t} t^{\alpha-1} dt \right] \\ &= \lim_{\delta \rightarrow 0} \left(e^{-\delta} \delta^{\alpha} - e^{-(\delta^{-1})} \delta^{-\alpha} + \alpha \int_{\delta}^{\delta^{-1}} e^{-t} t^{\alpha-1} dt \right) = \alpha \Gamma(\alpha)\end{aligned}$$

Now it is defined that $0! = 1$ and so $\Gamma(1) = 0!$. Suppose that $\Gamma(n+1) = n!$, what of $\Gamma(n+2)$? Is it $(n+1)!$? if so, then by induction, the proposition is established. From what was just shown,

$$\Gamma(n+2) = \Gamma(n+1)(n+1) = n!(n+1) = (n+1)!$$

and so this proves the proposition. ■

The properties of the gamma function also allow for a fairly easy proof about differentiating under the integral in a Laplace transform. First is a definition.

Definition 10.2.5 A function ϕ has exponential growth on $[0, \infty)$ if there are positive constants λ, C such that $|\phi(t)| \leq Ce^{\lambda t}$ for all $t \geq 0$.

Theorem 10.2.6 Let $f(s) = \int_0^{\infty} e^{-st} \phi(t) dt$ where $t \rightarrow \phi(t) e^{-st}$ is improper Riemann integrable for all s large enough and ϕ has exponential growth. Then for s large enough, $f^{(k)}(s)$ exists and equals $\int_0^{\infty} (-t)^k e^{-st} \phi(t) dt$.

Proof: Suppose true for some $k \geq 0$. By definition it is so for $k = 0$. Then always assuming $s > \lambda, |h| < s - \lambda$, where $|\phi(t)| \leq Ce^{\lambda t}, \lambda \geq 0$,

$$\begin{aligned}\frac{f^{(k)}(s+h) - f^{(k)}(s)}{h} &= \int_0^{\infty} (-t)^k \frac{e^{-(s+h)t} - e^{-st}}{h} \phi(t) dt \\ &= \int_0^{\infty} (-t)^k e^{-st} \left(\frac{e^{-ht} - 1}{h} \right) \phi(t) dt = \int_0^{\infty} (-t)^k e^{-st} \left((-t) e^{\theta(h,t)} \right) \phi(t) dt\end{aligned}$$

where $\theta(h, t)$ is between $-ht$ and 0, this by the mean value theorem. Thus by mean value theorem again,

$$\begin{aligned}&\left| \frac{f^{(k)}(s+h) - f^{(k)}(s)}{h} - \int_0^{\infty} (-t)^{k+1} e^{-st} \phi(t) dt \right| \\ &\leq \int_0^{\infty} |t|^{k+1} C e^{\lambda t} e^{-st} |e^{\theta(h,t)} - 1| dt \leq \int_0^{\infty} t^{k+1} C e^{\lambda t} e^{-st} e^{\alpha(h,t)} |ht| dt \\ &\leq \int_0^{\infty} t^{k+2} C e^{\lambda t} e^{-st} |h| e^{t|h|} dt = C|h| \int_0^{\infty} t^{k+2} e^{-(s-(\lambda+|h|))t} dt\end{aligned}$$

Let $u = (s - (\lambda + |h|))t, du = (s - (\lambda + |h|))dt$. Then the above equals

$$C|h| \int_0^{\infty} \left(\frac{u}{s - (\lambda + |h|)} \right)^{k+2} e^{-u} \frac{1}{(s - (\lambda + |h|))} du$$

$$= \frac{C|h|}{(s - (\lambda + |h|))^{k+3}} \int_0^\infty e^{-u} u^{k+2} du = \frac{C|h|}{(s - (\lambda + |h|))^{k+3}} \Gamma(k+3)$$

Thus, as $h \rightarrow 0$, this converges to 0 and so this proves the theorem. ■

The function $f(s)$ just defined is called the Laplace transform of ϕ .

10.3 Laplace Transforms

The Laplace transform is extremely useful in differential equations because it can change a differential equation into an algebraic equation. This easy equation can then be solved and then you go backwards in a table of Laplace transforms to find the solution to the differential equation. It is also useful in statistics, where it is called a moment generating function, and integral equations.

Suppose f is a piecewise continuous, bounded function, on each $[0, R]$. Then from Corollary 7.3.14 $t \rightarrow f(t)$ is integrable on $[0, R]$ for each $R > 0$. So is $t \rightarrow e^{-st} f(t)$.

Definition 10.3.1 We say that a function defined on $[0, \infty)$ has exponential growth if for some $\lambda \geq 0$, and $C > 0$, $|f(t)| \leq Ce^{\lambda t}$

Note that this condition is satisfied if $|f(t)| \leq a + be^{\lambda t}$. You simply pick $C > \max(a, b)$ and observe that $a + be^{\lambda t} \leq 2Ce^{\lambda t}$.

Proposition 10.3.2 Let f have exponential growth and be piecewise continuous on $[0, R]$ for each R . Then

$$\lim_{R \rightarrow \infty} \int_0^R f(t) e^{-st} dt \equiv \mathcal{L}f(s)$$

exists for every $s > \lambda$ where $|f(t)| \leq e^{\lambda t}$. That limit is denoted as

$$\int_0^\infty f(t) e^{-st} dt.$$

Proof: Let $R_n \rightarrow \infty$. Then for $R_m < R_n$,

$$\begin{aligned} \left| \int_0^{R_m} f(t) e^{-st} dt - \int_0^{R_n} f(t) e^{-st} dt \right| &\leq \int_{R_m}^{R_n} |f(t)| e^{-st} dt \\ &\leq \int_{R_m}^{R_n} e^{-(s-\lambda)t} dt \leq e^{-(s-\lambda)R_m} \end{aligned}$$

The elementary computations are left to the reader. Then this converges to 0 as $R_m \rightarrow \infty$. It follows that $\left\{ \int_0^{R_n} f(t) e^{-st} dt \right\}_{n=1}^\infty$ is a Cauchy sequence and so it converges to $I \in \mathbb{R}$. The above computation shows that if \hat{R}_n also converges to ∞ as $n \rightarrow \infty$, then

$$\lim_{n \rightarrow \infty} \int_0^{R_n} f(t) e^{-st} dt = \lim_{n \rightarrow \infty} \int_0^{\hat{R}_n} f(t) e^{-st} dt$$

and so the limit does indeed exist and this defines the improper integral $\int_0^\infty f(t) e^{-ts} dt$. ■

Certain properties are obvious. For example,

1. If a, b scalars and if g, f have exponential growth, then for all s large enough,

$$\mathcal{L}(af + bg)(s) = a\mathcal{L}(f)(s) + b\mathcal{L}(g)(s)$$

2. If $f'(t)$ exists and has exponential growth, and so does $f(t)$ then for s large enough,

$$\mathcal{L}(f')(s) = -f(0) + s\mathcal{L}(f)(s)$$

One can also compute Laplace transforms of many standard functions without much difficulty. That which is most certainly not obvious is the following major theorem. This is the thing which is omitted from virtually all ordinary differential equations books, and it is this very thing which justifies the use of Laplace transforms. Without it or something like it, the whole method is nonsense. I am following [32]. This theorem says that if you know the Laplace transform, this will determine the function it came from at every point of continuity of this function. The proof is fairly technical but only involves the theory of the integral which was presented in this chapter.

Theorem 10.3.3 *Let ϕ have exponential growth and have finitely many discontinuities on every interval $[0, R]$ and let $f(s) \equiv \mathcal{L}(\phi)(s)$. Then if t is a point of continuity of ϕ , it follows that*

$$\phi(t) = \lim_{k \rightarrow \infty} \frac{(-1)^k}{k!} \left[f^{(k)}\left(\frac{k}{t}\right) \right] \left(\frac{k}{t}\right)^{k+1}.$$

Thus $\phi(t)$ is determined by its Laplace transform at every point of continuity.

Proof: First note that for k a positive integer, you can change the variable letting $ku = t$ and obtain

$$\frac{k^{k+1}}{k!} \int_0^\infty (e^{-u}u)^k du = \frac{k^{k+1}}{k!} \int_0^\infty e^{-t} \left(\frac{t}{k}\right)^k \frac{1}{k} dt$$

The details involve doing this on finite intervals using the theory of the Riemann integral developed earlier and then passing to a limit. Thus the above equals

$$\frac{1}{k!} \int_0^\infty e^{-t} t^k dt = \Gamma(k+1) \frac{1}{k!} = k! \frac{1}{k!} = 1$$

by Proposition 10.2.4.

Now assuming that $|\phi(u)| \leq Ce^{\lambda u}$, then from what was just shown,

$$\frac{k^{k+1}}{k!} \int_0^\infty (e^{-u}u)^k \phi(u) du - \phi(1) = \int_0^\infty \frac{k^{k+1}}{k!} (e^{-u}u)^k (\phi(u) - \phi(1)) du$$

Assuming ϕ is continuous at 1, the improper integral is of the form

$$\begin{aligned} & \int_0^{1-\delta} \frac{k^{k+1}}{k!} (e^{-u}u)^k (\phi(u) - \phi(1)) du + \int_{1-\delta}^{1+\delta} \frac{k^{k+1}}{k!} (e^{-u}u)^k (\phi(u) - \phi(1)) du \\ & + \int_{1+\delta}^\infty \frac{k^{k+1}}{k!} (e^{-u}u)^k (\phi(u) - \phi(1)) du \end{aligned}$$

Consider the first integral in the above. Letting K be an upper bound for

$$|\phi(u) - \phi(1)|$$

on $[0, 1]$,

$$\left| \int_0^{1-\delta} \frac{k^{k+1}}{k!} (e^{-u}u)^k (\phi(u) - \phi(1)) du \right| \leq K \int_0^{1-\delta} \frac{k^{k+1}}{k!} (e^{-u}u)^k du$$

$$\leq K \frac{k^{k+1}}{k!} \left(e^{-(1-\delta)} (1-\delta) \right)^k (1-\delta)$$

Now this converges to 0 as $k \rightarrow \infty$. In fact, for $0 < a < 1$,

$$\lim_{k \rightarrow \infty} \frac{k^{k+1}}{k!} (e^{-a} a)^k = 0$$

To see this, take the ratio of the $k+1$ term to the k^{th} term.

$$\frac{\frac{(k+1)^{k+2}}{(k+1)!} (e^{-a} a)^{k+1}}{\frac{k^{k+1}}{k!} (e^{-a} a)^k} = e^{-a} a \frac{1}{k+1} \frac{(k+1)^{k+2}}{k^{k+1}} = e^{-a} a \left(1 + \frac{1}{k}\right)^k \left(1 + \frac{1}{k}\right)$$

which converges to $e^{1-a} a$, a positive number less than 1. Verify this. You can see it is true by graphing xe^{1-x} on $[0, 1]$ for example. Therefore, denoting as A_k the expression $\frac{k^{k+1}}{k!} (e^{-a} a)^k$, and letting $e^{1-a} a < r < 1$, it follows that for all k large enough,

$$\frac{A_{k+1}}{A_k} < r < 1$$

and so, iterating this,

$$\frac{A_{k+m}}{A_k} = \frac{A_{k+m}}{A_{k+m-1}} \frac{A_{k+m-1}}{A_{k+m-2}} \frac{A_{k+m-2}}{A_{k+m-3}} \dots \frac{A_{k+1}}{A_k} \leq r^{m-1}$$

Since $|r| < 1$, $\lim_{m \rightarrow \infty} A_{k+m} \leq \lim_{m \rightarrow \infty} A_k r^{m-1} = 0$. Here $a = 1 - \delta$.

Next consider the last integral. This obviously converges to 0 because of the exponential growth of ϕ . In fact,

$$\left| \int_{1+\delta}^{\infty} \frac{k^{k+1}}{k!} (e^{-u} u)^k (\phi(u) - \phi(1)) du \right| \leq \int_{1+\delta}^{\infty} \frac{k^{k+1}}{k!} (e^{-u} u)^k (a + be^{\lambda u}) du$$

Now changing the variable letting $uk = t$, and doing everything on finite intervals followed by passing to a limit, the absolute value of the above is dominated by

$$\begin{aligned} & \int_{k(1+\delta)}^{\infty} \frac{k^{k+1}}{k!} e^{-t} \left(\frac{t}{k}\right)^k \frac{1}{k} (a + be^{\lambda(t/k)}) dt \\ &= \int_{k(1+\delta)}^{\infty} \frac{1}{k!} e^{-t} t^k (a + be^{\lambda(t/k)}) dt \text{ for some } a, b \geq 0 \\ &= \int_0^{\infty} \frac{1}{k!} e^{-t} t^k (a + be^{\lambda(t/k)}) dt - \int_0^{k(1+\delta)} \frac{1}{k!} e^{-t} t^k (a + be^{\lambda(t/k)}) dt \end{aligned}$$

However, the limit as $k \rightarrow \infty$ of the integral on the right equals the improper integral on the left. Thus this converges to 0 as $k \rightarrow \infty$. Thus all that is left to consider is the middle integral in which δ was chosen such that $|\phi(u) - \phi(1)| < \varepsilon$. Then from what was shown earlier,

$$\left| \int_{1-\delta}^{1+\delta} \frac{k^{k+1}}{k!} (e^{-u} u)^k (\phi(u) - \phi(1)) du \right| \leq \varepsilon \int_0^{\infty} \frac{k^{k+1}}{k!} (e^{-u} u)^k du = \varepsilon$$

It follows that if ϕ is continuous at 1,

$$\lim_{k \rightarrow \infty} \int_0^\infty \frac{k^{k+1}}{k!} (e^{-u}u)^k (\phi(u) - \phi(1)) du = 0$$

and so $\int_0^\infty \frac{k^{k+1}}{k!} (e^{-u}u)^k \phi(u) du = \phi(1)$. Now you simply replace $\phi(u)$ with $\phi(tu)$ where ϕ is continuous at t . This function of u still has exponential growth and is continuous at $u = 1$. Thus we obtain

$$\lim_{k \rightarrow \infty} \int_0^\infty \frac{k^{k+1}}{k!} (e^{-u}u)^k \phi(tu) du = \phi(t)$$

Now use Theorem 10.2.6 on

$$f(s) \equiv \int_0^\infty e^{-st} \phi(t) dt$$

This theorem says that for large s , $f^{(k)}(s)$ exists and equals $\int_0^\infty (-u)^k e^{-su} \phi(u) du$. Then

$$\frac{(-1)^k}{k!} \left[f^{(k)} \left(\frac{k}{t} \right) \right] \left(\frac{k}{t} \right)^{k+1} = \frac{(-1)^k}{k!} \left[\int_0^\infty (-u)^k e^{-(k/t)u} \phi(u) du \right] \left(\frac{k}{t} \right)^{k+1}$$

Now letting $v = \frac{u}{t}$, this reduces to

$$\frac{(-1)^k}{k!} \left[\int_0^\infty (-tv)^k e^{-kv} \phi(tv) t dv \right] \left(\frac{k}{t} \right)^{k+1} = \frac{k^{k+1}}{k!} \int_0^\infty e^{-kv} v^k \phi(tv) dv$$

which was shown above to converge to $\phi(t)$. ■

I think the approach given above is really interesting because it gives an explicit description of $\phi(t)$ at every point. However, there are other ways to show this. See my single variable advanced calculus book for another approach based on the Weierstrass approximation theorem. However, to really do it right, one should use complex variable techniques. You can actually get the inverse Laplace transform from doing contour integrals. It is in my book on calculus of real and complex variables and in the single variable book just mentioned. This is called the Bromwich integral, another kind of improper integral and it converges to the mid point of the jump of the function. It or something like it is actually used by computer algebra systems to invert Laplace transforms.

10.4 Exercises

1. The improper integrals discussed in the chapter had to do with an infinite interval of integration. Another kind of improper integral is considered when you try to integrate an unbounded function. Here is an example:

$$\int_0^1 \frac{1}{\sqrt{x}} dx \equiv \lim_{\varepsilon \rightarrow 0} \int_\varepsilon^1 \frac{1}{\sqrt{x}} dx$$

Find $\int_0^1 \frac{1}{x^\alpha} dx$ for various values of α . Consider what happens when $\alpha < 1$ and when $\alpha \geq 1$.

2. When f is Riemann integrable on $[a, R]$ for each $R > a$ the “improper” integral is defined as follows. $\int_a^\infty f(t) dt \equiv \lim_{R \rightarrow \infty} \int_a^R f(t) dt$ whenever this limit exists. Show $\int_0^\infty \frac{\sin x}{x} dx$ exists. Here the integrand is defined to equal 1 when $x = 0$, not that this matters.
3. Show $\int_0^\infty \sin(t^2) dt$ exists.
4. Suppose f is a continuous function which is not equal to zero on $[0, b]$. Show that

$$\int_0^b \frac{f(x)}{f(x) + f(b-x)} dx = \frac{b}{2}.$$

Hint: First change the variables to obtain the integral equals

$$\int_{-b/2}^{b/2} \frac{f(y+b/2)}{f(y+b/2) + f(b/2-y)} dy$$

Next show by another change of variables that this integral equals

$$\int_{-b/2}^{b/2} \frac{f(b/2-y)}{f(y+b/2) + f(b/2-y)} dy.$$

Thus the sum of these equals b .

5. Letting $[a, b] = [-\pi, \pi]$, consider an example of a regular Sturm Liouville problem which is of the form

$$y'' + \lambda y = 0, y(-\pi) = 0, y(\pi) = 0.$$

Show that if $\lambda = n^2$ and $y_n(x) = \sin(nx)$ for n a positive integer, then y_n is a solution to this regular Sturm Liouville problem. In this case, $q(x) = 1$ and so from Problem 19, it must be the case that

$$\int_{-\pi}^{\pi} \sin(nx) \sin(mx) dx = 0$$

if $n \neq m$. Show directly using integration by parts that the above equation is true.

6. Let $f : [a, b] \times [c, d] \rightarrow \mathbb{R}$ satisfy the following condition at $(x_0, y_0) \in [a, b] \times [c, d]$. For every $\varepsilon > 0$ there exists a $\delta > 0$ possibly depending on (x_0, y_0) such that if

$$\max(|x - x_0|, |y - y_0|) < \delta$$

then

$$|f(x, y) - f(x_0, y_0)| < \varepsilon.$$

This is what it means for f to be continuous at (x_0, y_0) . Show that if f is continuous at every point of $[a, b] \times [c, d]$, then it is uniformly continuous on $[a, b] \times [c, d]$. That is, for every $\varepsilon > 0$ there exists a $\delta > 0$ such that if $(x_0, y_0), (x, y)$ are any two points of $[a, b] \times [c, d]$ such that

$$\max(|x - x_0|, |y - y_0|) < \delta,$$

then

$$|f(x, y) - f(x_0, y_0)| < \varepsilon.$$

Also show that such a function achieves its maximum and its minimum on $[a, b] \times [c, d]$. **Hint:** This is easy if you follow the same procedure that was used earlier but you take subsequences for each component to show $[a, b] \times [c, d]$ is sequentially compact.

7. Suppose f is a real valued function defined on $[a, b] \times [c, d]$ which is uniformly continuous as described in Problem 6 and bounded which follow from an assumption that it is continuous. Show

$$x \rightarrow \int_c^d f(x, y) dy, y \rightarrow \int_a^b f(x, y) dx$$

are both continuous functions. The idea is you fix one of the variables, x in the first and then integrate the continuous function of y obtaining a real number which depends on the value of x fixed. Explain why it makes sense to write

$$\int_a^b \int_c^d f(x, y) dy dx, \int_c^d \int_a^b f(x, y) dx dy.$$

Now consider the first of the above iterated integrals. (That is what these are called.) Consider the following argument in which you fill in the details.

$$\begin{aligned} \int_a^b \int_c^d f(x, y) dy dx &= \sum_{i=1}^n \int_{x_{i-1}}^{x_i} \int_c^d f(x, y) dy dx \\ &= \sum_{i=1}^n \int_{x_{i-1}}^{x_i} \sum_{j=1}^m \int_{y_{j-1}}^{y_j} f(x, y) dy dx = \sum_{i=1}^n \sum_{j=1}^m \int_{x_{i-1}}^{x_i} \int_{y_{j-1}}^{y_j} f(x, y) dy dx \\ &= \sum_{i=1}^n \sum_{j=1}^m \int_{x_{i-1}}^{x_i} (y_j - y_{j-1}) f(x, t_j) dx \\ &= \sum_{i=1}^n \sum_{j=1}^m (y_j - y_{j-1}) (x_i - x_{i-1}) f(s_i, t_j) \end{aligned}$$

Also

$$\int_c^d \int_a^b f(x, y) dx dy = \sum_{j=1}^m \sum_{i=1}^n (y_j - y_{j-1}) (x_i - x_{i-1}) f(s'_i, t'_j)$$

and now because of uniform continuity, it follows that if the partition points are close enough,

$$|f(s'_j, t'_j) - f(s_j, t_j)| < \frac{\varepsilon}{(d-c)(b-a)}$$

and so

$$\left| \int_c^d \int_a^b f(x, y) dx dy - \int_a^b \int_c^d f(x, y) dy dx \right| < \varepsilon$$

Since ε is arbitrary, this shows the two iterated integrals are equal. This is a case of Fubini's theorem.

8. This problem is in Apostol [2]. Explain why whenever f is continuous on $[a, b]$

$$\lim_{n \rightarrow \infty} \frac{b-a}{n} \sum_{k=1}^n f\left(a + k\left(\frac{b-a}{n}\right)\right) = \int_a^b f dx.$$

Apply this to $f(x) = \frac{1}{1+x^2}$ on the interval $[0, 1]$ to obtain the very interesting formula $\frac{\pi}{4} = \lim_{n \rightarrow \infty} \sum_{k=1}^n \frac{n}{n^2+k^2}$.

9. Suppose $f : [a, b] \times (c, d) \rightarrow \mathbb{R}$ is continuous. This means that if $t_n \rightarrow t$ in $[a, b]$ and $x_n \rightarrow x$ in (c, d) , then $\lim_{n \rightarrow \infty} f(t_n, x_n) = f(t, x)$. Partial derivatives involve fixing one variable and taking the derivative with respect to the other. Thus the partial derivative of f with respect to the second variable, denoted as $\frac{\partial f}{\partial x}(t, x)$ is given by

$$\frac{\partial f}{\partial x}(t, x) \equiv \lim_{h \rightarrow 0} \frac{f(t, x+h) - f(t, x)}{h}$$

Suppose also $x \rightarrow \frac{\partial f}{\partial x}(t, x)$ exists and is continuous and that for some K independent of t ,

$$\left| \frac{\partial f}{\partial x}(t, z) - \frac{\partial f}{\partial x}(t, x) \right| < K |z - x|.$$

This last condition happens, for example if $\frac{\partial^2 f(t, x)}{\partial x^2}$ is uniformly bounded on $[a, b] \times (c, d)$. (Why?) Define $F(x) \equiv \int_a^b f(t, x) dt$. Take the difference quotient of F and show using the mean value theorem and the above assumptions that

$$F'(x) = \int_a^b \frac{\partial f(t, x)}{\partial x} dt.$$

Note that the above condition automatically implies $x \rightarrow \frac{\partial f}{\partial x}(t, x)$ is continuous.

10. This problem is on $\int_0^\infty e^{-x^2} dx$. First explain why the integral exists. Supply details in the following argument.

$$\begin{aligned} F(x) &\equiv \left(\int_0^x e^{-t^2} dt \right)^2, \quad F'(x) = 2 \left(\int_0^x e^{-t^2} dt \right) e^{-x^2} \\ &= 2x \left(\int_0^1 e^{-x^2 u^2} du \right) e^{-x^2}, \quad F(0) = 0 \end{aligned}$$

Then using Problem 7,

$$\begin{aligned} F(x) &= \int_0^x 2y \left(\int_0^1 e^{-y^2 u^2} du \right) e^{-y^2} dy = \int_0^1 \int_0^x 2y e^{-y^2(1+u^2)} dy du \\ &= \int_0^1 \left(\frac{1}{u^2+1} - \frac{e^{-x^2(u^2+1)}}{u^2+1} \right) du \end{aligned}$$

By uniform convergence considerations, (explain)

$$\left(\int_0^\infty e^{-t^2} dt \right)^2 = \int_0^1 \frac{1}{u^2+1} du = \arctan(1) = \frac{\pi}{4}.$$

11. Find $\Gamma\left(\frac{1}{2}\right)$. **Hint:** $\Gamma\left(\frac{1}{2}\right) \equiv \int_0^\infty e^{-t} t^{-1/2} dt$. Explain carefully why this equals

$$2 \int_0^\infty e^{-u^2} du$$

Then use Problem 10. Find a formula for $\Gamma\left(\frac{3}{2}\right), \Gamma\left(\frac{5}{2}\right)$, etc.

12. Verify that $\mathcal{L}(\sin(\omega t)) = \frac{\omega}{\omega^2 + s^2}$ and $\mathcal{L}(\cos(t)) = \frac{s}{\omega^2 + s^2}$.
13. It was shown that $\mathcal{L}(\sin(t)) = \frac{1}{1+s^2}$. Show that it makes sense to take $\mathcal{L}\left(\frac{\sin t}{t}\right)$. Show that

$$\int_0^\infty \frac{\sin(t)}{t} e^{-st} dt = \frac{\pi}{2} - \int_0^s \frac{1}{1+u^2} du \quad (*)$$

To do this, let $f(s) = \int_0^\infty \frac{\sin(t)}{t} e^{-st} dt$ and show using Theorem 10.2.6 that

$$f'(s) = -\frac{1}{1+s^2} \text{ so } f(s) = -\arctan(s) + C$$

Then, by changing variables, argue that as $s \rightarrow \infty, f(s) \rightarrow 0$. Use this to determine C . Then when you have done this, you will have an interesting formula valid for all positive s . To finish it, let $s = 0$. Assume f is continuous from the right at 0.

14. Show that $\mathcal{L}(y') = s\mathcal{L}(y) - y(0)$. Then explain why $\mathcal{L}(y'') = s^2\mathcal{L}(y) - sy(0) - y'(0)$. Give a general formula for $\mathcal{L}(y^{(k)})$ where $y^{(k)}$ denotes the k^{th} derivative.
15. Suppose you have the differential equation with initial condition

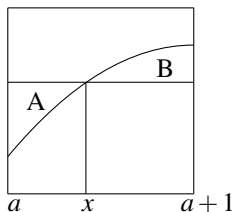
$$y'' + \omega^2 y = 0, y(0) = 1, y'(0) = 0$$

Use the above problem and the fact that $\mathcal{L}(\sin(\omega t)) = \frac{\omega}{\omega^2 + s^2}$ and $\mathcal{L}(\cos(t)) = \frac{s}{\omega^2 + s^2}$ to find the solution to this initial value problem. Solve the same problem with initial condition $y(0) = 0, y'(0) = 1$. Now give the solution to the differential equation with initial condition

$$y'' + \omega^2 y = 0, y(0) = a, y'(0) = b$$

Congratulations, you just found the general solution to the equation of undamped oscillation.

16. Use the mean value theorem for integrals, in Proposition 7.1.4 on Page 185 to conclude that $\int_a^{a+1} \ln(t) dt = \ln(x) \leq \ln\left(a + \frac{1}{2}\right)$ for some $x \in (a, a+1)$. **Hint:** Consider the shape of the graph of $\ln(x)$ in the following picture. Explain why if x is the special value between a and $a+1$, then the area of A is equal to area of B. Why should $x < a + \frac{1}{2}$?



Now use this to obtain the inequality 10.1.

17. Let r be a positive integer. Then if $f(x) = \frac{1}{\Gamma(r/2)2^{r/2}}x^{(r/2)-1}e^{-x/2}$, this function is called a chi-squared density, denoted as $\mathcal{X}^2(r)$. Show for each r , $\int_0^\infty f(x)dx = 1$. This particular function is the basis for a large part of mathematical statistics.
18. Suppose f is a continuous function and

$$\int_a^b f(x)x^n dx = 0$$

for $n = 0, 1, 2, 3, \dots$. Show that $f(x) = 0$ for all x . **Hint:** You might use the Weierstrass approximation theorem.

19. Suppose $|g(t)| < Ce^{-\delta t}$ for some $\delta > 0$, g continuous and defined for $t \geq 0$. Suppose also that whenever $s \geq 1$, $\int_0^\infty g(t)e^{-st}dt = 0$. Show that then $g(t) = 0$. **Hint:** Let $u = e^{-t}$. Then the integral reduces to $\int_0^1 u^{s-1}g(-\ln(u))du$. Here you define $\phi(u) = g(-\ln(u))$ if $u > 0$ and $\phi(0) = 0$. Then from the growth assumption, ϕ is continuous. Now use the previous problem.
20. Suppose f is continuous and $|f(t)| \leq Ce^{\lambda t}$, $\lambda > 0$ so it has exponential growth. Then suppose that if $s \geq s_0$, $\int_0^\infty e^{-st}f(t)dt = 0$. In other words, $\mathcal{L}(f(t))(s) = 0$ for all s large enough. Then consider $g(t) \equiv e^{-(\lambda+s_0+\delta)t}f(t)$. Then if $s \geq 1$, $\int_0^\infty e^{-st}g(t)dt = 0$. Hence $g(t) = 0 = f(t)$. Fill in the details.
21. To show you the power of Stirling's formula, find whether the series $\sum_{n=1}^\infty \frac{n!e^n}{n^n}$ converges. The ratio test falls flat but you can try it if you like. Now explain why, if n is large enough, $n! \geq \frac{1}{2}\sqrt{\pi}\sqrt{2}e^{-n}n^{n+(1/2)} \equiv c\sqrt{2}e^{-n}n^{n+(1/2)}$
22. Let f, g be continuous. Show that

$$\int_0^R \int_0^t f(t-u)g(u)dudt = \int_0^R \int_u^R f(t-u)g(u)dtdu.$$

Hint: The formula $\int_0^R \int_0^t f(t)g(u)dudt = \int_0^R \int_u^R f(t)g(u)dtdu$ is pretty easy. If f is a polynomial, then $f(t-u)$ is the sum of things like $c_k t^k u^{m-k}$. Then you could use the Weierstrass approximation theorem to get the general result.

23. If $F(s), G(s)$ are the Laplace transforms of $f(t), g(t)$ respectively, define $f * g(t) \equiv \int_0^t f(t-u)g(u)du$. Show the Laplace transform of $f * g$ is $F(s)G(s)$ and that if f, g have exponential growth, then so does $f * g$.
24. Verify the following short table of Laplace transforms. $f(t)$ denotes the function and $F(s)$ denotes its Laplace transform. **Hint:** You might use induction on some of these.

$f(t)$	$F(s)$	$f(t)$	$F(s)$	$f(t)$	$F(s)$
$t^n e^{at}$	$\frac{n!}{(s-a)^{n+1}}$	$t^n, n \in \mathbb{N}$	$\frac{n!}{s^{n+1}}$	$e^{at} \sin bt$	$\frac{b}{(s-a)^2 + b^2}$
$e^{at} \cos bt$	$\frac{s-a}{(s-a)^2 + b^2}$	$f * g(t)$	$F(s)G(s)$		

25. Maybe f has exponential growth and finitely many jumps in any finite interval, but $\int_0^\infty e^{-st}f(t)dt = 0$ for all s large enough. In this case, let $F(t) \equiv \int_0^t f(u)du$. Use integration by parts to verify that for all large enough s , $\int_0^R F(t)e^{-st}dt = \int_0^\infty \frac{e^{-st}}{s}f(t)dt = 0$. Therefore, by what was shown in the chapter, $F(t) = 0$. Now use the fundamental

theorem of calculus to conclude $f(t) = 0$ except for the jumps. Explain why if f, g are continuous except for finitely many jumps on each finite interval with exponential growth and same Laplace transform for large s , then $f = g$ except for jumps.

Chapter 11

Power Series

11.1 Functions Defined in Terms of Series

Earlier Taylor series expansions were discussed for given functions. More generally power series can be used to define new functions which you may not have a name for. This is actually the most exciting thing about power series, their ability to define new functions as a limit of polynomials.

Definition 11.1.1 Let $\{a_k\}_{k=0}^{\infty}$ be a sequence of numbers. The expression,

$$\sum_{k=0}^{\infty} a_k (x-a)^k \quad (11.1)$$

is called a Taylor series or power series centered at a . It is understood that x and $a \in \mathbb{R}$. More generally, these variables will be complex numbers, but in this book, only real numbers.

In the above definition, x is a variable. Thus you can put in various values of x and ask whether the resulting series of numbers converges. Defining D to be the set of all values of x such that the resulting series does converge, define a new function f defined on D having values in \mathbb{R} as

$$f(x) \equiv \sum_{k=0}^{\infty} a_k (x-a)^k.$$

This might be a totally new function, one which has no name. Nevertheless, much can be said about such functions. The following lemma is fundamental in considering the form of D which always turns out to be of the form $B(a, r)$ along with possibly some points z such that $|z-a| = r$. First here is a simple lemma which will be useful.

Lemma 11.1.2 $\lim_{n \rightarrow \infty} n^{1/n} = 1$.

Proof: It is clear $n^{1/n} \geq 1$. Let $n^{1/n} = 1 + e_n$ where $0 \leq e_n$. Then raising both sides to the n^{th} power for $n > 1$ and using the binomial theorem,

$$n = (1 + e_n)^n = \sum_{k=0}^n \binom{n}{k} e_n^k \geq 1 + ne_n + (n(n-1)/2) e_n^2 \geq (n(n-1)/2) e_n^2$$

Thus $0 \leq e_n^2 \leq \frac{n}{n(n-1)} = \frac{1}{n-1}$. From this the desired result follows because

$$\left| n^{1/n} - 1 \right| = e_n \leq \frac{1}{\sqrt{n-1}}. \blacksquare$$

Theorem 11.1.3 *Let $\sum_{k=0}^{\infty} a_k (x-a)^k$ be a Taylor series. Then there exists $r \leq \infty$ such that the Taylor series converges absolutely if $|x-a| < r$. Furthermore, if $|x-a| > r$, the Taylor series diverges. If $\lambda < r$ then the Taylor series converges uniformly on the closed disk $|x-a| \leq \lambda$.*

Proof: See Definition 3.3.16 for the notion of \limsup and \liminf . Note

$$\limsup_{k \rightarrow \infty} \left| a_k (x-a)^k \right|^{1/k} = \limsup_{k \rightarrow \infty} |a_k|^{1/k} |x-a|.$$

Then by the root test, the series converges absolutely if

$$|x-a| \limsup_{k \rightarrow \infty} |a_k|^{1/k} < 1$$

and diverges if

$$|x-a| \limsup_{k \rightarrow \infty} |a_k|^{1/k} > 1.$$

Thus define

$$r \equiv \begin{cases} 1 / \limsup_{k \rightarrow \infty} |a_k|^{1/k} & \text{if } \infty > \limsup_{k \rightarrow \infty} |a_k|^{1/k} > 0 \\ \infty & \text{if } \limsup_{k \rightarrow \infty} |a_k|^{1/k} = 0 \\ 0 & \text{if } \limsup_{k \rightarrow \infty} |a_k|^{1/k} = \infty \end{cases}$$

Next let λ be as described. Then if $|x-a| \leq \lambda$, then

$$\limsup_{k \rightarrow \infty} \left| a_k (x-a)^k \right|^{1/k} = \limsup_{k \rightarrow \infty} |a_k|^{1/k} |x-a| \leq \lambda \limsup_{k \rightarrow \infty} |a_k|^{1/k} \leq \frac{\lambda}{r} < \alpha < 1$$

It follows that for all k large enough and such x , $|a_k (x-a)^k| < \alpha^k$. Then by the Weierstrass M test, convergence is uniform. \blacksquare

Note that the radius of convergence r is given by

$$\limsup_{k \rightarrow \infty} |a_k|^{1/k} r = 1$$

Definition 11.1.4 *The number in the above theorem is called the radius of convergence and the set on which convergence takes place is called the disc of convergence. Since this book only considers functions of one real variable, it will be called the interval of convergence.*

Now the theorem was proved using the root test but often you use the ratio test to find the interval of convergence. This kind of thing is typical in math so get used to it. The proof of a theorem does not always yield a way to find the thing the theorem speaks about. The above is an existence theorem. There exists an interval of convergence from the above theorem. You find it in specific cases any way that is most convenient.

Example 11.1.5 Find the interval of convergence of the Taylor series $\sum_{n=1}^{\infty} \frac{x^n}{n}$.

Use Corollary 6.3.3.

$$\lim_{n \rightarrow \infty} \left(\frac{|x|^n}{n} \right)^{1/n} = \lim_{n \rightarrow \infty} \frac{|x|}{\sqrt[n]{n}} = |x|$$

because $\lim_{n \rightarrow \infty} \sqrt[n]{n} = 1$ and so if $|x| < 1$ the series converges. The points satisfying $|z| = 1$ require special attention. When $x = 1$ the series diverges because it reduces to $\sum_{n=1}^{\infty} \frac{1}{n}$. At $x = -1$ the series converges because it reduces to $\sum_{n=1}^{\infty} \frac{(-1)^n}{n}$ and the alternating series test applies and gives convergence.

What of the other numbers z satisfying $|z| = 1$? These numbers will be complex so outside the content of this book, but it turns out this series will converge at all these numbers by a use of the Dirichlet test.

Example 11.1.6 Find the radius of convergence of $\sum_{n=1}^{\infty} \frac{n^n}{n!} x^n$.

Apply the ratio test. Taking the ratio of the absolute values of the $(n+1)^{th}$ and the n^{th} terms

$$\frac{\frac{(n+1)^{(n+1)}}{(n+1)n!} |x|^{n+1}}{\frac{n^n}{n!} |x|^n} = (n+1)^n |x| n^{-n} = |x| \left(1 + \frac{1}{n} \right)^n \rightarrow |x|e$$

Therefore the series converges absolutely if $|x|e < 1$ and diverges if $|x|e > 1$. Consequently, $r = 1/e$ because

$$\lim_{n \rightarrow \infty} \left(1 + \frac{1}{n} \right)^n = e$$

To see this is the case, the limit, if it exists, is the same as

$$\lim_{x \rightarrow 0} (1+x)^{1/x} = \lim_{x \rightarrow 0} e^{\frac{\ln(1+x)}{x}} = e^{\lim_{x \rightarrow 0} \frac{\ln(1+x)}{x}} = e$$

from an application of L'Hopital's rule.

11.2 Operations on Power Series

It is desirable to be able to differentiate and multiply power series. The following theorem says you can differentiate power series in the most natural way on the interval of convergence, just as you would differentiate a polynomial. This theorem may seem obvious, but it is a serious mistake to think this. You usually cannot differentiate an infinite series whose terms are functions even if the functions are themselves polynomials. The following is special and pertains to power series. It is another example of the interchange of two limits, in this case, the limit involved in taking the derivative and the limit of the sequence of finite sums.

When you formally differentiate a series term by term, the result is called the derived series.

Theorem 11.2.1 Let $\sum_{n=0}^{\infty} a_n(x-a)^n$ be a Taylor series having radius of convergence $R > 0$ and let

$$f(x) \equiv \sum_{n=0}^{\infty} a_n(x-a)^n \quad (11.2)$$

for $|x - a| < R$. Then

$$f'(x) = \sum_{n=0}^{\infty} a_n n (x-a)^{n-1} = \sum_{n=1}^{\infty} a_n n (x-a)^{n-1} \quad (11.3)$$

and this new differentiated power series, the derived series, has radius of convergence equal to R .

Proof: First consider the claim that the derived series has radius of convergence equal to R . Let \hat{R} be the radius of convergence of the derived series. Then from Proposition 3.3.18 and Lemma 11.1.2,

$$\frac{1}{\hat{R}} \equiv \limsup_{n \rightarrow \infty} |a_n|^{1/n} n^{1/n} = \limsup_{n \rightarrow \infty} |a_n|^{1/n} \equiv \frac{1}{R},$$

so $\hat{R} = R$. If $\limsup_{n \rightarrow \infty} |a_n|^{1/n} = 0$, the same is true of $\limsup_{n \rightarrow \infty} |a_n|^{1/n} n^{1/n}$ and in this case, the series and derived series both have radius of convergence equal to ∞ .

Now let $r < R$, the radius of convergence of both series, and suppose $|x - a| < r$. Let δ be small enough that if $|h| < \delta$, then $|x + h - a| < r$ also. Then for $|h| < \delta$, consider the difference quotient.

$$\frac{f(x+h) - f(x)}{h} = \frac{1}{h} \sum_{k=0}^{\infty} a_k \left((x+h-a)^k - (x-a)^k \right)$$

By the mean value theorem, there exists $\theta_{kh} \in (0, 1)$ such that

$$\begin{aligned} \frac{f(x+h) - f(x)}{h} &= \frac{1}{h} \sum_{k=0}^{\infty} a_k \left((x+h-a)^k - (x-a)^k \right) \\ &= \frac{1}{h} \sum_{k=1}^{\infty} a_k k (x + \theta_{kh}h - a)^{k-1} h = \sum_{k=1}^{\infty} a_k k (x + \theta_{kh}h - a)^{k-1} \\ &= \sum_{k=1}^{\infty} a_k k \left[(x + \theta_{kh}h - a)^{k-1} - (x-a)^{k-1} \right] + \sum_{k=1}^{\infty} a_k k (x-a)^{k-1} \end{aligned}$$

By the mean value theorem again, there exists $\alpha_{kh} \in (0, 1)$ such that

$$= \sum_{k=2}^{\infty} \theta_{kh} h a_k k (k-1) (x + \alpha_{kh}h - a)^{k-2} + \sum_{k=1}^{\infty} a_k k (x-a)^{k-1}$$

The second series is the derived series. Consider the first.

$$\begin{aligned} \left| \sum_{k=2}^{\infty} \theta_{kh} h a_k k (k-1) (x + \alpha_{kh}h - a)^{k-2} \right| &\leq h \sum_{k=2}^{\infty} k(k-1) |a_k| |x + \alpha_{kh}h - a|^{k-2} \\ &\leq r^2 h \sum_{k=2}^{\infty} k(k-1) |a_k| r^k \end{aligned}$$

Now

$$\limsup_{k \rightarrow \infty} (k(k-1))^{1/k} |a_k|^{1/k} (r^k)^{1/k} = \limsup_{k \rightarrow \infty} |a_k|^{1/k} r = \frac{r}{R} < 1$$

and so the series converges by the root test. Hence, letting $h \rightarrow 0$ yields the desired result that

$$\lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} = \sum_{k=1}^{\infty} a_k k (x-a)^{k-1} \blacksquare$$

As an immediate corollary, it is possible to characterize the coefficients of a Taylor series.

Corollary 11.2.2 Let $\sum_{n=0}^{\infty} a_n (x-a)^n$ be a Taylor series with radius of convergence $r > 0$ and let

$$f(x) \equiv \sum_{n=0}^{\infty} a_n (x-a)^n. \quad (11.4)$$

Then

$$a_n = \frac{f^{(n)}(a)}{n!}. \quad (11.5)$$

Proof: From 11.4, $f(a) = a_0 \equiv f^{(0)}(a)/0!$. From Theorem 11.2.1,

$$f'(x) = \sum_{n=1}^{\infty} a_n n (x-a)^{n-1} = a_1 + \sum_{n=2}^{\infty} a_n n (x-a)^{n-1}.$$

Now let $x = a$ and obtain that $f'(a) = a_1 = f'(a)/1!$. Next use Theorem 11.2.1 again to take the second derivative and obtain

$$f''(x) = 2a_2 + \sum_{n=3}^{\infty} a_n n(n-1) (x-a)^{n-2}$$

let $x = a$ in this equation and obtain $a_2 = f''(a)/2 = f''(a)/2!$. Continuing this way proves the corollary. \blacksquare

This also shows the coefficients of a Taylor series are unique. That is, if

$$\sum_{k=0}^{\infty} a_k (x-a)^k = \sum_{k=0}^{\infty} b_k (x-a)^k$$

for all x in some open set containing a , then $a_k = b_k$ for all k .

Example 11.2.3 Find the sum $\sum_{k=1}^{\infty} k 2^{-k}$.

It may not be obvious what this sum equals but with the above theorem it is easy to find. From the formula for the sum of a geometric series, $\frac{1}{1-t} = \sum_{k=0}^{\infty} t^k$ if $|t| < 1$. Differentiate both sides to obtain

$$(1-t)^{-2} = \sum_{k=1}^{\infty} k t^{k-1}$$

whenever $|t| < 1$. Let $t = 1/2$. Then

$$4 = \frac{1}{(1-(1/2))^2} = \sum_{k=1}^{\infty} k 2^{-(k-1)}$$

and so if you multiply both sides by 2^{-1} , $2 = \sum_{k=1}^{\infty} k 2^{-k}$.

The above theorem shows that a power series is infinitely differentiable. Does it go the other way? That is, if the function has infinitely many continuous derivatives, is it correctly represented as a power series? The answer is no. See Problem 7 on Page 161 for an example. In fact, this is an important example and distinction. The modern theory of partial differential equations is built on just such functions which have many derivatives but no power series.

11.3 Power Series for Some Known Functions

If $x \rightarrow e^x$ has a power series, it is $\sum_{k=0}^{\infty} \frac{x^k}{k!}$. This is because of Corollary 11.2.2 and what was shown earlier that the derivative of this function is itself while $e^0 = 1$. Thus the question is whether this series really equals e^x . You can see that this is the case by looking at the Lagrange form of the remainder discussed earlier. In this case, it is

$$\frac{e^{\xi}}{(n+1)!} x^{n+1}$$

where ξ is some number between 0 and x . Does this remainder term converge to 0 as $n \rightarrow \infty$. The answer is yes because

$$\sum_{n=1}^{\infty} \frac{e^{\xi}}{(n+1)!} |x|^{n+1}$$

converges by the ratio test. Indeed,

$$\frac{\frac{e^{\xi}}{(n+2)!} |x|^{n+2}}{\frac{e^{\xi}}{(n+1)!} |x|^{n+1}} = |x| \frac{1}{n+2} \rightarrow 0$$

and by the n^{th} term test, it follows $\lim_{n \rightarrow \infty} \frac{e^{\xi}}{(n+1)!} x^{n+1} = 0$. Thus

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

Similar considerations show that

$$\sin(x) = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)!}, \quad \cos(x) = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k}}{(2k)!}$$

The details are left for you to do.

11.4 The Binomial Theorem

The following is a very important example known as the binomial series.

Example 11.4.1 Find a Taylor series for the function $(1+x)^{\alpha}$ centered at 0 valid for $|x| < 1$.

Use Theorem 11.2.1 to do this. First note that if $y(x) \equiv (1+x)^{\alpha}$, then y is a solution of the following initial value problem.

$$y' - \frac{\alpha}{(1+x)} y = 0, \quad y(0) = 1. \quad (11.6)$$

Next it is necessary to observe there is only one solution to this initial value problem. To see this, multiply both sides of the differential equation in 11.6 by $(1+x)^{-\alpha}$. When this is done, one obtains

$$\frac{d}{dx} ((1+x)^{-\alpha} y) = (1+x)^{-\alpha} \left(y' - \frac{\alpha}{(1+x)} y \right) = 0. \quad (11.7)$$

Therefore, from 11.7, there must exist a constant, C , such that

$$(1+x)^{-\alpha}y = C.$$

However, $y(0) = 1$ and so it must be that $C = 1$. Therefore, there is exactly one solution to the initial value problem in 11.6 and it is $y(x) = (1+x)^\alpha$.

The strategy for finding the Taylor series of this function consists of finding a series which solves the initial value problem above. Let

$$y(x) \equiv \sum_{n=0}^{\infty} a_n x^n \quad (11.8)$$

be a solution to 11.6. Of course it is not known at this time whether such a series exists. However, the process of finding it will demonstrate its existence. From Theorem 11.2.1 and the initial value problem,

$$(1+x) \sum_{n=0}^{\infty} a_n n x^{n-1} - \sum_{n=0}^{\infty} \alpha a_n x^n = 0$$

and so

$$\sum_{n=1}^{\infty} a_n n x^{n-1} + \sum_{n=0}^{\infty} a_n (n - \alpha) x^n = 0$$

Changing the variable of summation in the first sum,

$$\sum_{n=0}^{\infty} a_{n+1} (n+1) x^n + \sum_{n=0}^{\infty} a_n (n - \alpha) x^n = 0$$

and from Corollary 11.2.2 and the initial condition for 11.6 this requires

$$a_{n+1} = \frac{a_n (\alpha - n)}{n+1}, a_0 = 1. \quad (11.9)$$

Therefore, from 11.9 and letting $n = 0$, $a_1 = \alpha$, then using 11.9 again along with this information,

$$a_2 = \frac{\alpha(\alpha-1)}{2}.$$

Using the same process,

$$a_3 = \frac{\left(\frac{\alpha(\alpha-1)}{2}\right)(\alpha-2)}{3} = \frac{\alpha(\alpha-1)(\alpha-2)}{3!}.$$

By now you can spot the pattern. In general,

$$a_n = \frac{\overbrace{\alpha(\alpha-1) \cdots (\alpha-n+1)}^{n \text{ of these factors}}}{n!}.$$

Therefore, the candidate for the Taylor series is

$$y(x) = \sum_{n=0}^{\infty} \frac{\alpha(\alpha-1) \cdots (\alpha-n+1)}{n!} x^n.$$

Furthermore, the above discussion shows this series solves the initial value problem on its interval of convergence. It only remains to show the radius of convergence of this series equals 1. It will then follow that this series equals $(1+x)^\alpha$ because of uniqueness of the initial value problem. To find the radius of convergence, use the ratio test. Thus the ratio of the absolute values of $(n+1)^{\text{st}}$ term to the absolute value of the n^{th} term is

$$\frac{\left| \frac{\alpha(\alpha-1)\cdots(\alpha-n+1)(\alpha-n)}{(n+1)n!} \right| |x|^{n+1}}{\left| \frac{\alpha(\alpha-1)\cdots(\alpha-n+1)}{n!} \right| |x|^n} = |x| \frac{|\alpha-n|}{n+1} \rightarrow |x|$$

showing that the radius of convergence is 1 since the series converges if $|x| < 1$ and diverges if $|x| > 1$.

The expression, $\frac{\alpha(\alpha-1)\cdots(\alpha-n+1)}{n!}$ is often denoted as $\binom{\alpha}{n}$. With this notation, the following theorem has been established.

Theorem 11.4.2 *Let α be a real number and let $|x| < 1$. Then*

$$(1+x)^\alpha = \sum_{n=0}^{\infty} \binom{\alpha}{n} x^n.$$

There is a very interesting issue related to the above theorem which illustrates the limitation of power series. The function $f(x) = (1+x)^\alpha$ makes sense for all $x > -1$ but one is only able to describe it with a power series on the interval $(-1, 1)$. Think about this. The above technique is a standard one for obtaining solutions of differential equations and this example illustrates a deficiency in the method.

To completely understand power series, it is necessary to take a course in complex analysis. It turns out that the right way to consider Taylor series is through the use of geometric series and something called the Cauchy integral formula of complex analysis. However, these are topics for another course.

11.5 Exercises

1. Verify the power series claimed in the chapter for $\cos(x)$, $\sin(x)$ and e^x . The method for doing this was shown in the chapter in the case of e^x . Go through the details carefully and then do the same details for $\cos(x)$, $\sin(x)$.
2. The logarithm test states the following. Suppose $a_k \neq 0$ for large k and that $p = \lim_{k \rightarrow \infty} \frac{\ln\left(\frac{1}{|a_k|}\right)}{\ln k}$ exists. If $p > 1$, then $\sum_{k=1}^{\infty} a_k$ converges absolutely. If $p < 1$, then the series, $\sum_{k=1}^{\infty} a_k$ does not converge absolutely. Prove this theorem.
3. Using the Cauchy condensation test, determine the convergence of $\sum_{k=2}^{\infty} \frac{1}{k \ln k}$. Now determine the convergence of $\sum_{k=2}^{\infty} \frac{1}{k(\ln k)^{1.001}}$.
4. Find the values of p for which the following series converges and the values of p for which it diverges.

$$\sum_{k=4}^{\infty} \frac{1}{\ln^p(\ln(k)) \ln(k) k}$$

5. For p a positive number, determine the convergence of

$$\sum_{n=2}^{\infty} \frac{\ln n}{n^p}$$

for various values of p .

6. Suppose $\sum_{n=0}^{\infty} a_n (x-c)^n$ is a power series with radius of convergence r . Show the series converge uniformly on any interval $[a, b]$ where $[a, b] \subseteq (c-r, c+r)$. This is in the text but go through the details yourself.
7. In this problem, x will be a complex number. Thus you will find the disk of convergence, not just an interval of convergence. In other words, you will find all complex numbers such that the given series converges. Find the disc of convergence of the series $\sum \frac{x^n}{n^p}$ for various values of p . **Hint:** Use Dirichlet's test.
8. The power series for e^x was given above. Thus

$$e = \sum_{k=0}^{\infty} \frac{1}{k!}.$$

Show e is irrational. **Hint:** If $e = p/q$ for p, q positive integers, then argue

$$q! \left(\frac{p}{q} - \sum_{k=0}^q \frac{1}{k!} \right)$$

is an integer. However, you can also show

$$q! \left(\sum_{k=0}^{\infty} \frac{1}{k!} - \sum_{k=0}^q \frac{1}{k!} \right) < 1$$

9. Let $a \geq 1$. Show that for all $x > 0$, you have the inequality

$$ax > \ln(1+x^a).$$

10. Show

$$\frac{1}{1+x^2} = \sum_{k=0}^n (-1)^k x^{2k} + \frac{(-1)^{n+1} x^{2n+2}}{1+x^2}.$$

Now use this to find a series which converges to $\arctan(1) = \pi/4$. Recall

$$\arctan(x) = \int_0^x \frac{1}{1+t^2} dt.$$

For which values of x will your series converge? For which values of x does the above description of \arctan in terms of an integral make sense? Does this help to show the inferiority of power series?

11. Show

$$\arcsin(x) = \int_0^x \frac{1}{\sqrt{1-t^2}} dt.$$

Now use the binomial theorem to find a power series for $\arcsin(x)$.

11.6 Multiplication of Power Series

Next consider the problem of multiplying two power series.

Theorem 11.6.1 *Let $\sum_{n=0}^{\infty} a_n (x-a)^n$ and $\sum_{n=0}^{\infty} b_n (x-a)^n$ be two power series having radii of convergence r_1 and r_2 , both positive. Then*

$$\left(\sum_{n=0}^{\infty} a_n (x-a)^n \right) \left(\sum_{n=0}^{\infty} b_n (x-a)^n \right) = \sum_{n=0}^{\infty} \left(\sum_{k=0}^n a_k b_{n-k} \right) (x-a)^n$$

whenever $|x-a| < r \equiv \min(r_1, r_2)$.

Proof: By Theorem 11.1.3 both series converge absolutely if $|x-a| < r$. Therefore, by Theorem 6.6.7

$$\begin{aligned} & \left(\sum_{n=0}^{\infty} a_n (x-a)^n \right) \left(\sum_{n=0}^{\infty} b_n (x-a)^n \right) = \\ & \sum_{n=0}^{\infty} \sum_{k=0}^n a_k (x-a)^k b_{n-k} (x-a)^{n-k} = \sum_{n=0}^{\infty} \left(\sum_{k=0}^n a_k b_{n-k} \right) (x-a)^n. \blacksquare \end{aligned}$$

The significance of this theorem in terms of applications is that it states you can multiply power series just as you would multiply polynomials and everything will be all right on the common interval of convergence.

This theorem can be used to find Taylor series which would perhaps be hard to find without it. Here is an example.

Example 11.6.2 *Find the Taylor series for $e^x \sin x$ centered at $x = 0$.*

All that is required is to multiply

$$\left(\overbrace{1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} \cdots}^{e^x} \right) \left(\overbrace{x - \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots}^{\sin x} \right)$$

From the above theorem the result should be

$$\begin{aligned} & x + x^2 + \left(-\frac{1}{3!} + \frac{1}{2!} \right) x^3 + \cdots \\ & = x + x^2 + \frac{1}{3} x^3 + \cdots \end{aligned}$$

You can continue this way and get the following to a few more terms.

$$x + x^2 + \frac{1}{3} x^3 - \frac{1}{30} x^5 - \frac{1}{90} x^6 - \frac{1}{630} x^7 + \cdots$$

I don't see a pattern in these coefficients but I can go on generating them as long as I want. (In practice this tends to not be very long.) I also know the resulting power series will converge for all x because both the series for e^x and the one for $\sin x$ converge for all x .

Example 11.6.3 Find the Taylor series for $\tan x$ centered at $x = 0$.

Lets suppose it has a Taylor series $a_0 + a_1x + a_2x^2 + \dots$. Then

$$(a_0 + a_1x + a_2x^2 + \dots) \left(\overbrace{1 - \frac{x^2}{2} + \frac{x^4}{4!} + \dots}^{\cos x} \right) = \left(x - \frac{x^3}{3!} + \frac{x^5}{5!} + \dots \right).$$

Using the above, $a_0 = 0, a_1x = x$ so $a_1 = 1, (0 - \frac{1}{2})x^2 = 0$ so $a_2 = 0$. $(a_3 - \frac{a_1}{2})x^3 = \frac{-1}{3!}x^3$ so $a_3 - \frac{1}{2} = -\frac{1}{6}$ so $a_3 = \frac{1}{3}$. Clearly one can continue in this manner. Thus the first several terms of the power series for \tan are

$$\tan x = x + \frac{1}{3}x^3 + \dots$$

You can go on calculating these terms and find the next two yielding

$$\tan x = x + \frac{1}{3}x^3 + \frac{2}{15}x^5 + \frac{17}{315}x^7 + \dots$$

This is a very significant technique because, as you see, there does not appear to be a very simple pattern for the coefficients of the power series for $\tan x$. Of course there are some issues here about whether $\tan x$ even has a power series, but if it does, the above must be it. In fact, $\tan(x)$ will have a power series valid on some interval centered at 0 and this becomes completely obvious when one uses methods from complex analysis but it isn't too obvious at this point. If you are interested in this issue, read the last section of the chapter. Note also that what has been accomplished is to divide the power series for $\sin x$ by the power series for $\cos x$ just like they were polynomials.

11.7 Exercises

1. Find the radius of convergence of the following.

(a) $\sum_{k=1}^{\infty} \left(\frac{x}{2}\right)^n$

(d) $\sum_{n=0}^{\infty} \frac{(3n)^n}{(3n)!} x^n$

(b) $\sum_{k=1}^{\infty} \sin\left(\frac{1}{n}\right) 3^n x^n$

(e) $\sum_{n=0}^{\infty} \frac{(2n)^n}{(2n)!} x^n$

(c) $\sum_{k=0}^{\infty} k! x^k$

2. Find $\sum_{k=1}^{\infty} k 2^{-k}$.

4. Find $\sum_{k=1}^{\infty} \frac{2^{-k}}{k}$.

3. Find $\sum_{k=1}^{\infty} k^2 3^{-k}$.

5. Find $\sum_{k=1}^{\infty} \frac{3^{-k}}{k}$.

6. Find the power series centered at 0 for the function $1/(1+x^2)$ and give the radius of convergence. Where does the function make sense? Where does the power series equal the function?

7. Find a power series for the function $f(x) \equiv \frac{\sin(\sqrt{x})}{\sqrt{x}}$ for $x > 0$. Where does $f(x)$ make sense? Where does the power series you found converge?

8. Use the power series technique which was applied in Example 11.4.1 to consider the initial value problem $y' = y, y(0) = 1$. This yields another way to obtain the power series for e^x .
9. Use the power series technique on the initial value problem $y' + y = 0, y(0) = 1$. What is the solution to this initial value problem?
10. Use the power series technique to find solutions in terms of power series to the initial value problem

$$y'' + xy = 0, y(0) = 0, y'(0) = 1.$$

Tell where your solution gives a valid description of a solution for the initial value problem. **Hint:** This is a little different but you proceed the same way as in Example 11.4.1. The main difference is you have to do two differentiations of the power series instead of one.

11. Find several terms of a likely power series solution to the nonlinear initial value problem

$$y'' + a \sin(y) = 0, y(0) = 1, y'(0) = 0.$$

This is the equation which governs the vibration of a pendulum.

12. Suppose the function e^x is defined in terms of a power series, $e^x \equiv \sum_{k=0}^{\infty} \frac{x^k}{k!}$. Use Theorem 6.6.7 on Page 176 to show directly the usual law of exponents,

$$e^{x+y} = e^x e^y.$$

Be sure to check all the hypotheses.

13. Let $f_n(x) \equiv \left(\frac{1}{n} + x^2\right)^{1/2}$. Show that for all x ,

$$||x| - f_n(x)| \leq \frac{1}{\sqrt{n}}.$$

Thus these approximate functions converge uniformly to the function $f(x) = |x|$. Now show $f'_n(0) = 0$ for all n and so $f'_n(0) \rightarrow 0$. However, the function $f(x) \equiv |x|$ has no derivative at $x = 0$. Thus even though $f_n(x) \rightarrow f(x)$ for all x , you cannot say that $f'_n(0) \rightarrow f'(0)$.

14. Let the functions, $f_n(x)$ be given in Problem 13 and consider

$$g_1(x) = f_1(x), g_n(x) = f_n(x) - f_{n-1}(x) \text{ if } n > 1.$$

Show that for all $x, \sum_{k=0}^{\infty} g_k(x) = |x|$ and that $g'_k(0) = 0$ for all k . Therefore, you can't differentiate the series term by term and get the right answer¹.

15. Use the theorem about the binomial series to give a proof of the binomial theorem

$$(a+b)^n = \sum_{k=0}^n \binom{n}{k} a^{n-k} b^k$$

whenever n is a positive integer.

¹How bad can this get? It can be much worse than this. In fact, there are functions which are continuous everywhere and differentiable nowhere. We typically don't have names for them but they are there just the same. Every such function can be written as an infinite sum of polynomials which of course have derivatives at every point. Thus it is nonsense to differentiate an infinite sum term by term without a theorem of some sort.

16. Find the power series for $\sin(x^2)$ by plugging in x^2 where ever there is an x in the power series for $\sin x$. How do you know this is the power series for $\sin(x^2)$?
17. Find the first several terms of the power series for $\sin^2(x)$ by multiplying the power series for $\sin(x)$. Next use the trig. identity, $\sin^2(x) = \frac{1 - \cos(2x)}{2}$ and the power series for $\cos(2x)$ to find the power series.
18. Find the power series for $f(x) = \frac{1}{\sqrt{1-x^2}}$.
19. Let a, b be two positive numbers and let $p > 1$. Choose q such that

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Now verify the important inequality

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}.$$

Hint: You might try considering $f(a) = \frac{a^p}{p} + \frac{b^q}{q} - ab$ for fixed $b > 0$ and examine its graph using the derivative.

20. Using Problem 19, show that if $\alpha > 0, p > 1$, it follows that for all $x > 0$

$$\left(\frac{p-1}{p}x + \frac{\alpha}{p}x^{1-p} \right)^p \geq \alpha.$$

21. Using Problem 20, define for $p > 1$ and $\alpha > 0$ the following sequence

$$x_{n+1} \equiv \frac{p-1}{p}x_n + \frac{\alpha}{p}x_n^{1-p}, \quad x_1 > 0.$$

Show $\lim_{n \rightarrow \infty} x_n = x$ where $x = \alpha^{1/p}$. In fact show that after x_1 the sequence decreases to $\alpha^{1/p}$.

22. Recall that for a power series, $\sum_{k=0}^{\infty} a_k (x-c)^k$ you could differentiate term by term on the interval of convergence. Show that if the radius of convergence of the above series is $r > 0$ and if $[a, b] \subseteq (c-r, c+r)$, then

$$\begin{aligned} & \int_a^b \sum_{k=0}^{\infty} a_k (x-c)^k dx \\ &= a_0(b-a) + \sum_{k=1}^{\infty} \frac{a_k}{k} (b-c)^{k+1} - \sum_{k=1}^{\infty} \frac{a_k}{k} (a-c)^{k+1} \end{aligned}$$

In other words, you can integrate term by term.

11.8 Some Other Theorems

First recall Theorem 6.6.7 on Page 176. For convenience, the version of this theorem which is of interest here is listed below.

Theorem 11.8.1 Suppose $\sum_{i=0}^{\infty} a_i$ and $\sum_{j=0}^{\infty} b_j$ both converge absolutely. Then

$$\left(\sum_{i=0}^{\infty} a_i \right) \left(\sum_{j=0}^{\infty} b_j \right) = \sum_{n=0}^{\infty} c_n$$

where $c_n = \sum_{k=0}^n a_k b_{n-k}$. Furthermore, $\sum_{n=0}^{\infty} c_n$ converges absolutely.

Proof: It only remains to verify the last series converges absolutely. Letting p_{nk} equal 1 if $k \leq n$ and 0 if $k > n$. Then by Theorem 6.6.4 on Page 175

$$\begin{aligned} \sum_{n=0}^{\infty} |c_n| &= \sum_{n=0}^{\infty} \left| \sum_{k=0}^n a_k b_{n-k} \right| \leq \sum_{n=0}^{\infty} \sum_{k=0}^n |a_k| |b_{n-k}| = \sum_{n=0}^{\infty} \sum_{k=0}^{\infty} p_{nk} |a_k| |b_{n-k}| \\ &= \sum_{k=0}^{\infty} \sum_{n=0}^{\infty} p_{nk} |a_k| |b_{n-k}| = \sum_{k=0}^{\infty} \sum_{n=k}^{\infty} |a_k| |b_{n-k}| = \sum_{k=0}^{\infty} |a_k| \sum_{n=0}^{\infty} |b_n| < \infty. \blacksquare \end{aligned}$$

The above theorem is about multiplying two series. What if you wanted to consider $(\sum_{n=0}^{\infty} a_n)^p$ where p is a positive integer maybe larger than 2? Is there a similar theorem to the above?

Definition 11.8.2 Define

$$\sum_{k_1 + \dots + k_p = m} a_{k_1} a_{k_2} \cdots a_{k_p}$$

as follows. Consider all ordered lists of nonnegative integers k_1, \dots, k_p which have the property that $\sum_{i=1}^p k_i = m$. For each such list of integers, form the product, $a_{k_1} a_{k_2} \cdots a_{k_p}$ and then add all these products.

Note that $\sum_{k=0}^n a_k a_{n-k} = \sum_{k_1+k_2=n} a_{k_1} a_{k_2}$. Therefore, from the above theorem, if $\sum a_i$ converges absolutely, it follows

$$\left(\sum_{i=0}^{\infty} a_i \right)^2 = \sum_{n=0}^{\infty} \left(\sum_{k_1+k_2=n} a_{k_1} a_{k_2} \right).$$

It turns out a similar theorem holds for replacing 2 with p .

Theorem 11.8.3 Suppose $\sum_{n=0}^{\infty} a_n$ converges absolutely. Then if p is a positive integer,

$$\left(\sum_{n=0}^{\infty} a_n \right)^p = \sum_{m=0}^{\infty} c_{mp}$$

where

$$c_{mp} \equiv \sum_{k_1 + \dots + k_p = m} a_{k_1} \cdots a_{k_p}.$$

Proof: First note this is obviously true if $p = 1$ and is also true if $p = 2$ from the above theorem. Now suppose this is true for p and consider $(\sum_{n=0}^{\infty} a_n)^{p+1}$. By the induction

hypothesis and the above theorem on the Cauchy product,

$$\begin{aligned}
 \left(\sum_{n=0}^{\infty} a_n \right)^{p+1} &= \left(\sum_{n=0}^{\infty} a_n \right)^p \left(\sum_{n=0}^{\infty} a_n \right) = \left(\sum_{m=0}^{\infty} c_{mp} \right) \left(\sum_{n=0}^{\infty} a_n \right) \\
 &= \sum_{n=0}^{\infty} \left(\sum_{k=0}^n c_{kp} a_{n-k} \right) = \sum_{n=0}^{\infty} \sum_{k=0}^n \sum_{k_1+\dots+k_p=k} a_{k_1} \cdots a_{k_p} a_{n-k} \\
 &= \sum_{n=0}^{\infty} \sum_{k_1+\dots+k_{p+1}=n} a_{k_1} \cdots a_{k_{p+1}} \blacksquare
 \end{aligned}$$

This theorem implies the following corollary for power series.

Corollary 11.8.4 *Let*

$$\sum_{n=0}^{\infty} a_n (x-a)^n$$

be a power series having radius of convergence, $r > 0$. Then if $|x-a| < r$,

$$\left(\sum_{n=0}^{\infty} a_n (x-a)^n \right)^p = \sum_{n=0}^{\infty} b_{np} (x-a)^n$$

where

$$b_{np} \equiv \sum_{k_1+\dots+k_p=n} a_{k_1} \cdots a_{k_p}.$$

Proof: Since $|x-a| < r$, the series, $\sum_{n=0}^{\infty} a_n (x-a)^n$, converges absolutely. Therefore, the above theorem applies and

$$\begin{aligned}
 \left(\sum_{n=0}^{\infty} a_n (x-a)^n \right)^p &= \sum_{n=0}^{\infty} \left(\sum_{k_1+\dots+k_p=n} a_{k_1} (x-a)^{k_1} \cdots a_{k_p} (x-a)^{k_p} \right) \\
 &= \sum_{n=0}^{\infty} \left(\sum_{k_1+\dots+k_p=n} a_{k_1} \cdots a_{k_p} \right) (x-a)^n. \blacksquare
 \end{aligned}$$

With this theorem it is possible to consider the question raised in Example 11.6.3 on Page 263 about the existence of the power series for $\tan x$. This question is clearly included in the more general question of when $(\sum_{n=0}^{\infty} a_n (x-a)^n)^{-1}$ has a power series.

Lemma 11.8.5 *Let $f(x) = \sum_{n=0}^{\infty} a_n (x-a)^n$, a power series having radius of convergence $r > 0$. Suppose also that $f(a) = 1$. Then there exists $r_1 > 0$ and $\{b_n\}$ such that for all $|x-a| < r_1$, $\frac{1}{f(x)} = \sum_{n=0}^{\infty} b_n (x-a)^n$.*

Proof: By continuity, there exists $r_1 > 0$ such that if $|x-a| < r_1$, then

$$\sum_{n=1}^{\infty} |a_n| |x-a|^n < 1.$$

Now pick such an x . Then

$$\frac{1}{f(x)} = \frac{1}{1 + \sum_{n=1}^{\infty} a_n (x-a)^n} = \frac{1}{1 + \sum_{n=0}^{\infty} c_n (x-a)^n}$$

where $c_n = a_n$ if $n > 0$ and $c_0 = 0$. Then

$$\left| \sum_{n=1}^{\infty} a_n (x-a)^n \right| \leq \sum_{n=1}^{\infty} |a_n| |x-a|^n < 1 \quad (11.10)$$

and so from the formula for the sum of a geometric series,

$$\frac{1}{f(x)} = \sum_{p=0}^{\infty} \left(- \sum_{n=0}^{\infty} c_n (x-a)^n \right)^p.$$

By Corollary 11.8.4, this equals

$$\sum_{p=0}^{\infty} \sum_{n=0}^{\infty} b_{np} (x-a)^n \quad (11.11)$$

where $b_{np} = \sum_{k_1+\dots+k_p=n} (-1)^p c_{k_1} \cdots c_{k_p}$. Thus

$$|b_{np}| \leq \sum_{k_1+\dots+k_p=n} |c_{k_1}| \cdots |c_{k_p}| \equiv B_{np}$$

and so by Theorem 11.8.3,

$$\sum_{p=0}^{\infty} \sum_{n=0}^{\infty} |b_{np}| |x-a|^n \leq \sum_{p=0}^{\infty} \sum_{n=0}^{\infty} B_{np} |x-a|^n = \sum_{p=0}^{\infty} \left(\sum_{n=0}^{\infty} |c_n| |x-a|^n \right)^p < \infty$$

by 11.10 and the formula for the sum of a geometric series. Since the series of 11.11 converges absolutely, Theorem 6.6.4 on Page 175 implies the series in 11.11 equals

$$\sum_{n=0}^{\infty} \left(\sum_{p=0}^{\infty} b_{np} \right) (x-a)^n$$

and so, letting $\sum_{p=0}^{\infty} b_{np} \equiv b_n$, this proves the lemma. ■

With this lemma, the following theorem is easy to obtain.

Theorem 11.8.6 *Let $f(x) = \sum_{n=0}^{\infty} a_n (x-a)^n$, a power series having radius of convergence $r > 0$. Suppose also that $f(a) \neq 0$. Then there exists $r_1 > 0$ and $\{b_n\}$ such that for all $|x-a| < r_1$, $\frac{1}{f(x)} = \sum_{n=0}^{\infty} b_n (x-a)^n$.*

Proof: Let $g(x) \equiv f(x)/f(a)$ so that $g(x)$ satisfies the conditions of the above lemma. Then by that lemma, there exists $r_1 > 0$ and a sequence, $\{b_n\}$ such that

$$\frac{f(a)}{f(x)} = \sum_{n=0}^{\infty} b_n (x-a)^n$$

for all $|x-a| < r_1$. Then $\frac{1}{f(x)} = \sum_{n=0}^{\infty} \tilde{b}_n (x-a)^n$ where $\tilde{b}_n = b_n/f(a)$. ■

There is a very interesting question related to r_1 in this theorem. Consider $f(x) = 1+x^2$. In this case $r = \infty$ but the power series for $1/f(x)$ converges only if $|x| < 1$. What happens is this, $1/f(x)$ will have a power series that will converge for $|x-a| < r_1$ where r_1 is the distance between a and the nearest singularity or zero of $f(x)$ in the complex plane. In the case of $f(x) = 1+x^2$ this function has a zero at $x = \pm i$. This is just another instance of why the natural setting for the study of power series is the complex plane. To read more on power series, you should see the book by Apostol [3] or any text on complex variable. The best way to understand power series is to use methods of complex analysis.

11.9 Some Historical Observations

As mentioned earlier, one of the ill defined notions in calculus was the infinitesimal, dx . What is it? No one knew what exactly it was. It wasn't any positive real number and it wasn't 0 either. However, people thought in terms of $\frac{dy}{dx}$ and this was the derivative so they wished to understand the quotient of these unknown things. Gradually it became clear that whatever meaning the quotient had, it was closely connected to the methods for finding it and these methods eventually became the definition of its meaning, being formalized as the concept of limit. This was done by Bolzano early in 1800's.

Even though the notion of dx was not very well defined, the notation turned out to be very useful as in the methods presented above for changing variables in an integral.

The concept of an integral also developed gradually. It was possible to consider most of the physical applications in terms of an initial value problem for an unknown function y satisfying $y'(x) = f(x)$, $y(0) = y_0$ and this is essentially what was done in the 1700's, but this did not resolve fundamental questions concerning the existence of the integral. Of course this was impossible without a careful definition of what was meant by the integral which did not exist at that time. These kinds of questions were not considered very much in the 1700's and were first addressed by Cauchy around 1823 who considered what we call one sided Riemann sums for continuous functions. Since such a definition gives the integral for continuous functions, Cauchy's proof of the fundamental theorem of calculus was the first one which was complete although it is not clear whether he had all the details regarding uniform continuity, a concept developed later by Weierstrass. It is unsatisfactory to prove a theorem about something you have not defined precisely and before Cauchy, this was the state of the fundamental theorem of calculus. Riemann's improved description of the integral dates from around 1854 and was completed later by Darboux who proved the theorem about his integral and the Riemann integral being equivalent.

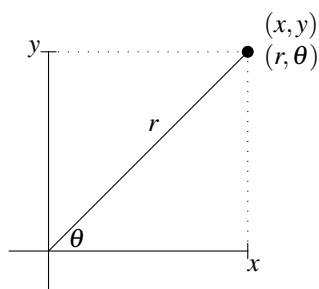
Newton discovered the binomial theorem for $(1+x)^\alpha$ in 1665. It is certainly a marvelous thing, but the importance of this and other power series tended to be over emphasized for much of the 1700's. Power series became much more understandable with the invention and development of complex analysis. This subject was continually expanded during the 1800's starting with Cauchy and continuing with most of the other mathematicians of that century.

What we now refer to as real analysis began early in the 1800's with the work of Bolzano. It was an effort to make calculus rigorous by removing intuitive geometric reasoning. Later on Weierstrass found nowhere differentiable continuous functions and Peano found examples of space filling continuous curves. Weierstrass also showed the importance of uniform convergence and uniform continuity. Eventually calculus was brought to its present form through his efforts. Of course the entire subject is built on completeness of \mathbb{R} . Dedekind and Cantor constructed \mathbb{R} from the rational numbers in 1872 although Dedekind did it earlier in 1858, but before this time, the mathematicians of that century used the essential characteristics of \mathbb{R} in their development of calculus. Dedekind, Cantor, and Weierstrass completed the removal of geometry from the foundations of calculus.

Chapter 12

Polar Coordinates

So far points have been identified in terms of Cartesian coordinates but there are other ways of specifying points in twodimensions. These other ways involve using a list of two or three numbers which have a totally different meaning than Cartesian coordinates to specify a point in two or three dimensional space. In general these lists of numbers which have a different meaning than Cartesian coordinates are called curvilinear coordinates. Probably the simplest curvilinear coordinate system is that of **polar coordinates**. The idea is suggested in the following picture.



You see in this picture, the number r identifies the distance of the point from the origin, $(0,0)$ while θ is the angle shown between the positive x axis and the line from the origin to the point. This angle will always be given in radians and is in the interval $[0, 2\pi)$. Thus the given point, indicated by a small dot in the picture, can be described in terms of the Cartesian coordinates (x,y) or the polar coordinates (r,θ) . How are the two coordinates systems related? From the picture,

$$x = r \cos(\theta), y = r \sin(\theta). \quad (12.1)$$

Example 12.0.1 The polar coordinates of a point in the plane are $(5, \frac{\pi}{6})$. Find the Cartesian or rectangular coordinates of this point.

From 12.1, $x = 5 \cos(\frac{\pi}{6}) = \frac{5}{2}\sqrt{3}$ and $y = 5 \sin(\frac{\pi}{6}) = \frac{5}{2}$. Thus the Cartesian coordinates are $(\frac{5}{2}\sqrt{3}, \frac{5}{2})$.

Example 12.0.2 Suppose the Cartesian coordinates of a point are $(3,4)$. Find the polar coordinates.

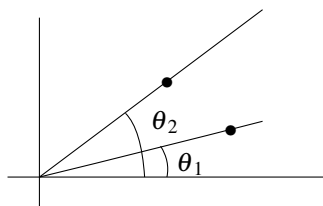
Recall that r is the distance from $(0, 0)$ and so $r = 5 = \sqrt{3^2 + 4^2}$. It remains to identify the angle. Note the point is in the first quadrant. (Both the x and y values are positive.) Therefore, the angle is something between 0 and $\pi/2$ and also $3 = 5 \cos(\theta)$, and $4 = 5 \sin(\theta)$. Therefore, dividing yields $\tan(\theta) = 4/3$. At this point, use a calculator or a table of trigonometric functions to find that at least approximately, $\theta = .927295$ radians.

12.1 Graphs in Polar Coordinates

Just as in the case of rectangular coordinates, it is possible to use relations between the polar coordinates to specify points in the plane. The process of sketching their graphs is very similar to that used to sketch graphs of functions in rectangular coordinates. I will only consider the case where the relation between the polar coordinates is of the form, $r = f(\theta)$. To graph such a relation, you can make a table of the form

θ	r
θ_1	$f(\theta_1)$
θ_2	$f(\theta_2)$
\vdots	\vdots

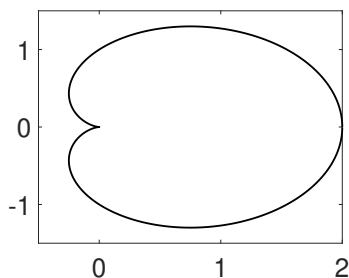
and then graph the resulting points and connect them up with a curve. The following picture illustrates how to begin this process.



To obtain the point in the plane which goes with the pair $(\theta, f(\theta))$, you draw the ray through the origin which makes an angle of θ with the positive x axis. Then you move along this ray a distance of $f(\theta)$ to obtain the point. As in the case with rectangular coordinates, this process is tedious and is best done by a computer algebra system.

Example 12.1.1 Graph the polar equation $r = 1 + \cos \theta$.

Using a computer algebra system, here is the graph of this cardioid.

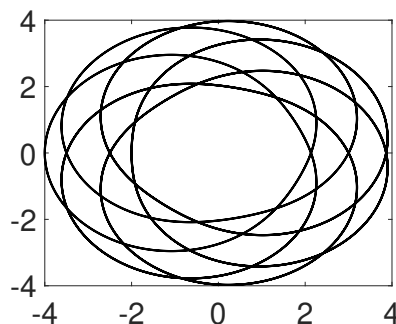


You can also see just from your knowledge of the trig. functions that the graph should look something like this. When $\theta = 0, r = 2$ and then as θ increases to $\pi/2$, you see

that $\cos \theta$ decreases to 0. Thus the line from the origin to the point on the curve should get shorter as θ goes from 0 to $\pi/2$. Then from $\pi/2$ to π , $\cos \theta$ gets negative eventually equaling -1 at $\theta = \pi$. Thus $r = 0$ at this point. Viewing the graph, you see this is exactly what happens. The above function is called a **cardioid**.

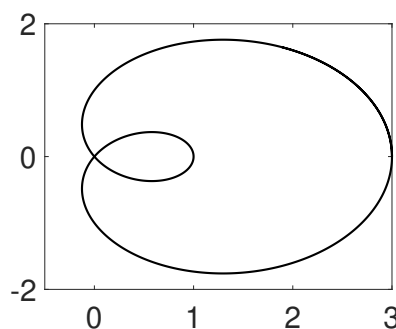
Here is another example. This is the graph obtained from $r = 3 + \sin\left(\frac{7\theta}{6}\right)$.

Example 12.1.2 Graph $r = 3 + \sin\left(\frac{7\theta}{6}\right)$ for $\theta \in [0, 14\pi]$.



In polar coordinates people sometimes allow r to be negative. When this happens, it means that to obtain the point in the plane, you go in the opposite direction along the ray which starts at the origin and makes an angle of θ with the positive x axis. I do not believe the fussiness occasioned by this extra generality is justified by any sufficiently interesting application so no more will be said about this. It is mainly a fun way to obtain pretty pictures. Here is such an example.

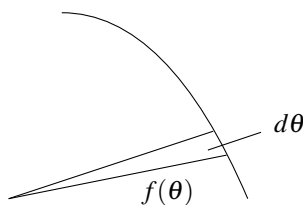
Example 12.1.3 Graph $r = 1 + 2\cos \theta$ for $\theta \in [0, 2\pi]$.



12.2 The Area in Polar Coordinates

How can you find the area of the region determined by $0 \leq r \leq f(\theta)$ for $\theta \in [a, b]$, assuming this is a well defined set of points in the plane? See Example 12.1.3 with $\theta \in [0, 2\pi]$ to see something which it would be better to avoid.

I have in mind the situation where every ray through the origin having angle θ for $\theta \in [a, b]$, $b - a \leq 2\pi$, intersects the graph of $r = f(\theta)$ in exactly one point. To see how to find the area of such a region, consider the following picture.



This is a representation of a small triangle obtained from two rays whose angles differ by only $d\theta$. What is the area of this triangle, dA ? It would be

$$\frac{1}{2} \sin(d\theta) f(\theta)^2 \approx \frac{1}{2} f(\theta)^2 d\theta = dA$$

with the approximation getting better as the angle gets smaller. Thus the area should solve the initial value problem,

$$\frac{dA}{d\theta} = \frac{1}{2} f(\theta)^2, \quad A(a) = 0.$$

Therefore, the total area would be given by the integral

$$\frac{1}{2} \int_a^b f(\theta)^2 d\theta. \quad (12.2)$$

Example 12.2.1 Find the area of the cardioid, $r = 1 + \cos \theta$ for $\theta \in [0, 2\pi]$.

From the graph of the cardioid presented earlier, you can see the region of interest satisfies the conditions above that every ray intersects the graph in only one point. Therefore, from 12.2 this area is

$$\frac{1}{2} \int_0^{2\pi} (1 + \cos(\theta))^2 d\theta = \frac{3}{2} \pi.$$

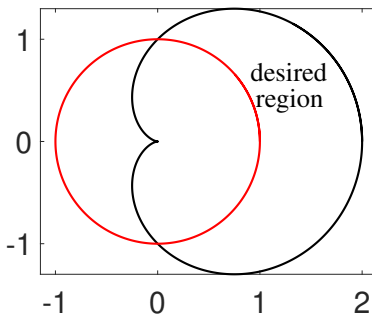
Example 12.2.2 Verify the area of a circle of radius a is πa^2 .

The polar equation is just $r = a$ for $\theta \in [0, 2\pi]$. Therefore, the area should be

$$\frac{1}{2} \int_0^{2\pi} a^2 d\theta = \pi a^2.$$

Example 12.2.3 Find the area of the region inside the cardioid, $r = 1 + \cos \theta$ and outside the circle $r = 1$ for $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$.

As is usual in such cases, it is a good idea to graph the curves involved to get an idea what is wanted.



The area of this region would be the area of the part of the cardioid corresponding to $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ minus the area of the part of the circle in the first quadrant. Thus the area is

$$\frac{1}{2} \int_{-\pi/2}^{\pi/2} (1 + \cos(\theta))^2 d\theta - \frac{1}{2} \int_{-\pi/2}^{\pi/2} 1 d\theta = \frac{1}{4}\pi + 2.$$

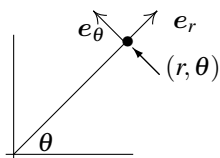
This example illustrates the following procedure for finding the area between the graphs of two curves given in polar coordinates.

Procedure 12.2.4 Suppose that for all $\theta \in [a, b]$, $0 < g(\theta) < f(\theta)$. To find the area of the region defined in terms of polar coordinates by $g(\theta) < r < f(\theta)$, $\theta \in [a, b]$, you do the following.

$$\frac{1}{2} \int_a^b (f(\theta)^2 - g(\theta)^2) d\theta.$$

12.3 The Acceleration in Polar Coordinates

I assume that by now, the reader has encountered Newton's laws of motion, especially the second law which gives the relationship, force equals mass times acceleration. Sometimes you have information about forces which act not in the direction of the coordinate axes but in some other direction. When this is the case, it is often useful to express things in terms of different coordinates which are consistent with these directions. A good example of this is the force exerted by the sun on a planet. This force is always directed toward the sun and so the force vector changes as the planet moves. To discuss this, consider the following simple diagram in which two unit vectors e_r and e_θ are shown.



The vector $e_r = (\cos \theta, \sin \theta)$ and the vector $e_\theta = (-\sin \theta, \cos \theta)$. Note that $e_\theta \cdot e_r = 0$. You should convince yourself that the directions of these two perpendicular vectors correspond to what is shown in the above picture. To help with this, note that $e_r \times e_\theta = \mathbf{k}$ if these vectors are considered as $e_\theta = (-\sin \theta, \cos \theta, 0)$, $e_r = (\cos \theta, \sin \theta, 0)$ and so $(e_r, e_\theta, \mathbf{k})$ forms a right hand system, so if you see that e_r points away from the origin, then it follows that e_θ points in the direction shown.

These two vectors also have the following relationship

$$e_\theta = \frac{de_r}{d\theta}, \quad e_r = -\frac{de_\theta}{d\theta}. \quad (12.3)$$

Now consider the position vector from $\mathbf{0}$ of a point in the plane, $\mathbf{r}(t)$. Then if $r(t)$, $\theta(t)$ are its polar coordinates at time t ,

$$\mathbf{r}(t) = r(t) e_r(\theta(t))$$

where $r(t) = |\mathbf{r}(t)|$. Thus $r(t)$ is just the distance from the origin $\mathbf{0}$ to the point. What are the velocity and acceleration in terms of e_r and e_θ ? Using the chain rule,

$$\frac{de_r}{dt} = \frac{de_r}{d\theta} \theta'(t), \quad \frac{de_\theta}{dt} = \frac{de_\theta}{d\theta} \theta'(t)$$

and so from 12.3,

$$\frac{de_r}{dt} = \theta'(t) e_\theta, \quad \frac{de_\theta}{dt} = -\theta'(t) e_r \quad (12.4)$$

Using 12.4 as needed along with the product rule and the chain rule,

$$\begin{aligned} \mathbf{r}'(t) &= r'(t) e_r + r(t) \frac{d}{dt}(e_r(\theta(t))) \\ &= r'(t) e_r + r(t) \theta'(t) e_\theta. \end{aligned}$$

Next consider the acceleration.

$$\begin{aligned} \mathbf{r}''(t) &= r''(t) e_r + r'(t) \frac{de_r}{dt} + r'(t) \theta'(t) e_\theta + r(t) \theta''(t) e_\theta + r(t) \theta'(t) \frac{d}{dt}(e_\theta) \\ &= r''(t) e_r + 2r'(t) \theta'(t) e_\theta + r(t) \theta''(t) e_\theta + r(t) \theta'(t) (-e_r) \theta'(t) \\ &= (r''(t) - r(t) \theta'(t)^2) e_r + (2r'(t) \theta'(t) + r(t) \theta''(t)) e_\theta. \end{aligned} \quad (12.5)$$

This is a very profound formula. Consider the following examples.

Example 12.3.1 Suppose an object of mass m moves at a uniform speed v , around a circle of radius R . Find the force acting on the object.

By Newton's second law, the force acting on the object is $m\mathbf{r}''$. In this case, $r(t) = R$, a constant and since the speed is constant, $\theta'' = 0$. Therefore, the term in 12.5 corresponding to e_θ equals zero and $m\mathbf{r}'' = -R\theta'(t)^2 e_r$. The speed of the object is v and so it moves v/R radians in unit time. Thus $\theta'(t) = v/R$ and so

$$m\mathbf{r}'' = -mR \left(\frac{v}{R}\right)^2 e_r = -m \frac{v^2}{R} e_r.$$

This is the familiar formula for centripetal force from elementary physics, obtained as a very special case of 12.5.

Example 12.3.2 A platform rotates at a constant speed in the counter clockwise direction and an object of mass m moves from the center of the platform toward the edge at constant speed along a line fixed in the rotating platform. What forces act on this object?

Let v denote the constant speed of the object moving toward the edge of the platform. Then

$$r'(t) = v, \quad r''(t) = 0, \quad \theta''(t) = 0,$$

while $\theta'(t) = \omega$, a positive constant. From 12.5

$$m\mathbf{r}''(t) = -mr(t) \omega^2 e_r + 2mv\omega e_\theta.$$

Thus the object experiences centripetal force from the first term and also a funny force from the second term which is in the direction of rotation of the platform. You can observe this by experiment if you like. Go to a playground and have someone spin one of those merry go rounds while you ride it and move from the center toward the edge. The term $2mv\omega e_\theta$ is called the Coriolis force.

12.4 The Fundamental Theorem of Algebra

The fundamental theorem of algebra states that every non constant polynomial having coefficients in \mathbb{C} has a zero in \mathbb{C} . If \mathbb{C} is replaced by \mathbb{R} , this is not true because of the example, $x^2 + 1 = 0$. This theorem is a very remarkable result and notwithstanding its title, all the most straightforward proofs depend on either analysis or topology. It was first mostly proved by Gauss in 1797. The first complete proof was given by Argand in 1806. I will give an informal explanation of this theorem which shows why it is reasonable to believe in the fundamental theorem of algebra. This will also introduce the idea of parametric curves in the plane of which much more will be said later. First is the notion of parametric curves in the plane.

Definition 12.4.1 For t in some interval, consider functions $t \rightarrow x(t), t \rightarrow y(t)$. This is called a *parametric function* or a *vector valued function* because you could consider $t \rightarrow (x(t), y(t))$. Thus $(x(t), y(t))$ yields a point in the plane or vector and as t changes, this point (vector) might move around yielding a curve in the plane. The real number t is called the *parameter*.

The explanation for the fundamental theorem of algebra involves arguing that some parametric curve must contain $0 + i0$.

Theorem 12.4.2 Let $p(z) = a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0$ where each a_k is a complex number and $a_n \neq 0, n \geq 1$. Then there exists $w \in \mathbb{C}$ such that $p(w) = 0$.

Here is the informal explanation. Dividing by the leading coefficient a_n , there is no loss of generality in assuming that the polynomial is of the form

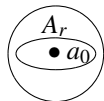
$$p(z) = z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0$$

If $a_0 = 0$, there is nothing to prove because $p(0) = 0$. Therefore, assume $a_0 \neq 0$. From the polar form of a complex number z , it can be written as $|z|(\cos \theta + i \sin \theta)$. Thus, by DeMoivre's theorem,

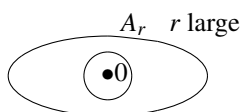
$$z^n = |z|^n (\cos(n\theta) + i \sin(n\theta))$$

It follows that z^n is some point on the circle of radius $|z|^n$

Denote by C_r the circle of radius r in the complex plane which is centered at 0. Then if r is sufficiently large and $|z| = r$, the term z^n is far larger than the rest of the polynomial. It is on the circle of radius $|z|^n$ while the other terms are on circles of fixed multiples of $|z|^k$ for $k \leq n-1$. Thus, for r large enough, $A_r = \{p(z) : z \in C_r\}$ describes a closed curve which misses the inside of some circle having 0 as its center. It won't be as simple as suggested in the following picture, but it will be a closed curve thanks to De Moivre's theorem and the observation that the cosine and sine are periodic. Now shrink r . Eventually, for r small enough, the non constant terms are negligible and so A_r is a curve which is contained in some circle centered at a_0 which has 0 on the outside.



r small



A_r r large

Thus it is reasonable to believe that for some r during this shrinking process, the set A_r must hit 0. It follows that $p(z) = 0$ for some z .

For example, consider the polynomial $x^3 + x + 1 + i$. It has no real zeros. However, you could let

$z = r(\cos t + i \sin t)$ and insert this into the polynomial. Thus you would want to find a point where

$$(r(\cos t + i \sin t))^3 + r(\cos t + i \sin t) + 1 + i = 0 + 0i$$

Expanding this expression on the left to write it in terms of real and imaginary parts, you get on the left

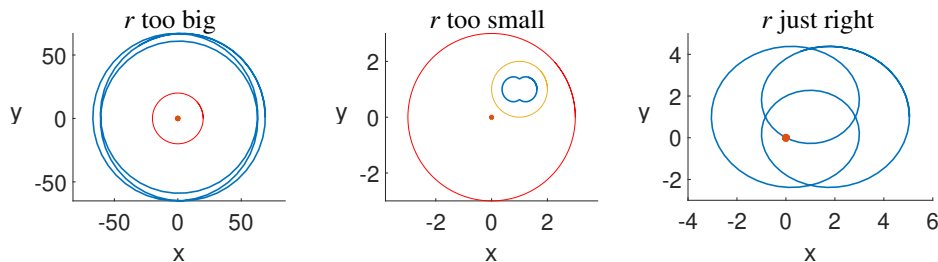
$$r^3 \cos^3 t - 3r^3 \cos t \sin^2 t + r \cos t + 1 + i(3r^3 \cos^2 t \sin t - r^3 \sin^3 t + r \sin t + 1)$$

Thus you need to have both the real and imaginary parts equal to 0. In other words, you need to have

$$r^3 \cos^3 t - 3r^3 \cos t \sin^2 t + r \cos t + 1 = 0$$

$$3r^3 \cos^2 t \sin t - r^3 \sin^3 t + r \sin t + 1 = 0$$

for some value of r and t . First here is a graph of this parametric function of t for $t \in [0, 2\pi]$ on the left, when $r = 4$. It is drawn by a computer and drawing it simply involves taking many values of t and connecting the resulting points with a curve just as you learned to graph in high school. Note how the graph misses the origin $0 + i0$. In fact, the closed curve surrounds a small circle which has the point $0 + i0$ on its inside.



Next is the graph when $r = .5$. Note how the closed curve is included in a circle which has $0 + i0$ on its outside. As you shrink r you get closed curves. At first, these closed curves enclose $0 + i0$ and later, they exclude $0 + i0$. Thus one of them should pass through this point. In fact, consider the curve which results when $r = 1.386$ which is the graph on the right. Note how for this value of r the curve passes through the point $0 + i0$. Thus for some t , $1.3862(\cos t + i \sin t)$ is a solution of the equation $p(z) = 0$.

Later I will give a real proof of this important theorem but this informal discussion shows why it is very reasonable to believe this theorem.

12.5 Polar Graphing in MATLAB

I think it is likely easiest to do these graphs by changing the equation in polar coordinates to one which is simply a parametric equation. Thus if you have

$$r = f(\theta), \theta \in [a, b]$$

You would write it parametrically as

$$x(\theta) = f(\theta) \cos(\theta), y(\theta) = f(\theta) \sin(\theta)$$

and you would graph the parametric curve

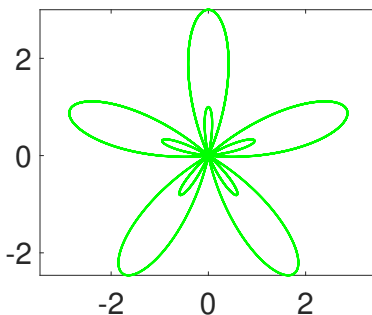
$$\theta \rightarrow (x(\theta), y(\theta))$$

Example 12.5.1 Graph the polar equation $r = 1 + 2 \sin(5\theta)$, $\theta \in [0, 7\pi]$. In general, it is much more fun to have MATLAB do the graphing for you.

Here is the syntax for this example. You will see how to modify it to make changes as desired. Try this and you can change line width and color as desired. To get to a new line, you press shift enter. To get MATLAB to do the graphing, you press enter. I used t as the parameter rather than θ because it is more convenient. In the top line, you are defining values of t which are .01 apart going from 0 to 7π . The reason you have $.$ * rather than simply * is that according to MATLAB, t is a list of numbers and you have to be doing something to the individual numbers in the list. If you wrote t^2 MATLAB would not know what you meant. It requires a little getting used to. However, if you use $\sin(t)$, it knows what is meant, so it seems to me it is not entirely consistent. Anyway, here is the syntax.

```
>>t=[0:.01:7*pi];
x=(1+2*sin(5*t)).*cos(t);
y=(1+2*sin(5*t)).*sin(t);
plot(x,y,'LineWidth',2,'color','green')
axis equal
```

When you do this, and press enter, you get



12.6 Exercises

1. Suppose $r = \frac{a}{1+\varepsilon \sin \theta}$ where $\varepsilon \geq 0$. By changing to rectangular coordinates, show that this is either a parabola, an ellipse or a hyperbola. Determine the values of ε which correspond to the various cases.
2. In Example 12.1.2 suppose you graphed it for $\theta \in [0, k\pi]$ where k is a positive integer. What is the smallest value of k such that the graph will start at $(3, 0)$ and end at $(3, 0)$?
3. Suppose you were to graph $r = 3 + \sin\left(\frac{m}{n}\theta\right)$ where m, n are integers. Can you give some description of what the graph will look like for $\theta \in [0, k\pi]$ for k a very large positive integer? How would things change if you did $r = 3 + \sin(\alpha\theta)$ where α is an irrational number?
4. Graph $r = 1 + \sin \theta$ for $\theta \in [0, 2\pi]$.
5. Graph $r = 2 + \sin \theta$ for $\theta \in [0, 2\pi]$.
6. Graph $r = 1 + 2 \sin \theta$ for $\theta \in [0, 2\pi]$.

7. Graph $r = 2 + \sin(2\theta)$ for $\theta \in [0, 2\pi]$.
8. Graph $r = 1 + \sin(2\theta)$ for $\theta \in [0, 2\pi]$.
9. Graph $r = 1 + \sin(3\theta)$ for $\theta \in [0, 2\pi]$.
10. Graph $r = \sin(3\theta) + 2 + \cos(3\theta)$ for $\theta \in [0, 2\pi]$.
11. Find the area of the bounded region determined by $r = 1 + \sin(3\theta)$ for $\theta \in [0, 2\pi]$.
12. Find the area inside $r = 1 + \sin \theta$ and outside the circle $r = 1/2$.
13. Find the area inside the circle $r = 1/2$ and outside the region defined by $r = 1 + \sin \theta$.

Chapter 13

Algebra and Geometry of \mathbb{R}^p

13.1 \mathbb{R}^p

The notation, \mathbb{R}^p refers to the collection of ordered lists of p numbers. The order matters. Thus $(1, 2, 3) \neq (3, 1, 2)$.

Definition 13.1.1 *Define*

$$\mathbb{R}^p \equiv \{(x_1, \dots, x_p) : x_j \in \mathbb{R} \text{ for } j = 1, \dots, p\}.$$

$(x_1, \dots, x_p) = (y_1, \dots, y_p)$ if and only if for all $j = 1, \dots, p$, $x_j = y_j$. When

$$(x_1, \dots, x_p) \in \mathbb{R}^p,$$

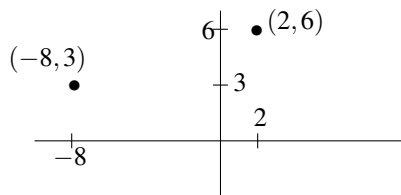
it is conventional to denote (x_1, \dots, x_p) by the single bold face letter \mathbf{x} . The numbers x_j are called the **coordinates**. The set

$$\{(0, \dots, 0, t, 0, \dots, 0) : t \in \mathbb{R}\}$$

for t in the i^{th} slot is called the i^{th} coordinate axis **coordinate axis**, the x_i axis for short. The point $\mathbf{0} \equiv (0, \dots, 0)$ is called the **origin**. Points in \mathbb{R}^p are also called **vectors**.

Thus $(1, 2, 4) \in \mathbb{R}^3$ and $(2, 1, 4) \in \mathbb{R}^3$ but $(1, 2, 4) \neq (2, 1, 4)$ because, even though the same numbers are involved, they don't match up. In particular, the first entries are not equal.

Why would anyone be interested in such a thing? First consider the case when $p = 1$. Then from the definition, $\mathbb{R}^1 = \mathbb{R}$. Recall that \mathbb{R} is identified with the points of a line. Look at the number line again. Observe that this amounts to identifying a point on this line with a real number. In other words a real number determines where you are on this line. Now suppose $p = 2$ and consider two lines which intersect each other at right angles as shown in the following picture.



Notice how you can identify a point shown in the plane with the ordered pair $(2, 6)$. You go to the right a distance of 2 and then up a distance of 6. Similarly, you can identify another point in the plane with the ordered pair $(-8, 3)$. Go to the left a distance of 8 and then up a distance of 3. The reason you go to the left is that there is a $-$ sign on the eight. From this reasoning, every ordered pair determines a unique point in the plane. Conversely, taking a point in the plane, you could draw two lines through the point, one vertical and the other horizontal and determine unique points x_1 on the horizontal line in the above picture and x_2 on the vertical line in the above picture, such that the point of interest is identified with the ordered pair (x_1, x_2) . In short, points in the plane can be identified with ordered pairs similar to the way that points on the real line are identified with real numbers. Now suppose $p = 3$. As just explained, the first two coordinates determine a point in a plane. Letting the third component determine how far up or down you go, depending on whether this number is positive or negative, this determines a point in space. Thus, $(1, 4, -5)$ would mean to determine the point in the plane that goes with $(1, 4)$ and then to go below this plane a distance of 5 to obtain a unique point in space. You see that the ordered triples correspond to points in space just as the ordered pairs correspond to points in a plane and single real numbers correspond to points on a line.

You can't stop here and say that you are only interested in $p \leq 3$. What if you were interested in the motion of two objects? You would need three coordinates to describe where the first object is and you would need another three coordinates to describe where the other object is located. Therefore, you would need to be considering \mathbb{R}^6 . If the two objects moved around, you would need a time coordinate as well. As another example, consider a hot object which is cooling and suppose you want the temperature of this object. How many coordinates would be needed? You would need one for the temperature, three for the position of the point in the object and one more for the time. Thus you would need to be considering \mathbb{R}^5 . Many other examples can be given. Sometimes p is very large. This is often the case in applications to business when they are trying to maximize profit subject to constraints. It also occurs in numerical analysis when people try to solve hard problems on a computer.

There are other ways to identify points in space with three numbers but the one presented is the most basic. In this case, the coordinates are known as **Cartesian coordinates** after Descartes¹ who invented this idea in the first half of the seventeenth century. I will often not bother to draw a distinction between the point in p dimensional space and its Cartesian coordinates but there really is such a distinction.

¹René Descartes 1596-1650 is often credited with inventing analytic geometry although it seems the ideas were actually known much earlier. He was interested in many different subjects, physiology, chemistry, and physics being some of them. He also wrote a large book in which he tried to explain the book of Genesis scientifically. Descartes ended up dying in Sweden.

13.2 Algebra in \mathbb{R}^p

There are two algebraic operations done with points of \mathbb{R}^p . One is addition and the other is multiplication by numbers, called scalars. Yes, numbers = scalars.

Definition 13.2.1 *If $x \in \mathbb{R}^p$ and a is a number, also called a **scalar**, then $ax \in \mathbb{R}^p$ is defined by*

$$ax = a(x_1, \dots, x_p) \equiv (ax_1, \dots, ax_p). \quad (13.1)$$

*This is known as **scalar multiplication**. If $x, y \in \mathbb{R}^p$ then $x + y \in \mathbb{R}^p$ and is defined by*

$$(x_1 + y_1, \dots, x_p + y_p) \quad (13.2)$$

*An element of \mathbb{R}^p $x \equiv (x_1, \dots, x_p)$ is called a **vector**. The above definition is known as **vector addition**.*

With this definition, the algebraic properties satisfy the conclusions of the following theorem. The conclusions of this theorem are called the **vector space axioms**. There are many other examples.

Theorem 13.2.2 *For v, w vectors in \mathbb{R}^p and α, β scalars, (real numbers), the following hold.*

$$v + w = w + v, \quad (13.3)$$

the commutative law of addition,

$$(v + w) + z = v + (w + z), \quad (13.4)$$

the associative law for addition,

$$v + \mathbf{0} = v, \quad (13.5)$$

the existence of an additive identity

$$v + (-v) = \mathbf{0}, \quad (13.6)$$

the existence of an additive inverse, Also

$$\alpha(v + w) = \alpha v + \alpha w, \quad (13.7)$$

$$(\alpha + \beta)v = \alpha v + \beta v, \quad (13.8)$$

$$\alpha(\beta v) = \alpha\beta(v), \quad (13.9)$$

$$1v = v. \quad (13.10)$$

In the above $\mathbf{0} = (0, \dots, 0)$.

You should verify these properties all hold. For example, consider 13.7.

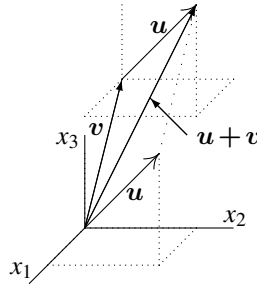
$$\begin{aligned} \alpha(v + w) &= \alpha(v_1 + w_1, \dots, v_p + w_p) = (\alpha(v_1 + w_1), \dots, \alpha(v_p + w_p)) \\ &= (\alpha v_1 + \alpha w_1, \dots, \alpha v_p + \alpha w_p) = (\alpha v_1, \dots, \alpha v_p) + (\alpha w_1, \dots, \alpha w_p) = \alpha v + \alpha w. \end{aligned}$$

As usual, subtraction is defined as $x - y \equiv x + (-y)$.

13.3 Geometric Meaning Of Vector Addition In \mathbb{R}^3

It was explained earlier that an element of \mathbb{R}^p is an p tuple of numbers and it was also shown that this can be used to determine a point in three dimensional space in the case where $p = 3$ and in two dimensional space, in the case where $p = 2$. This point was specified relative to some coordinate axes.

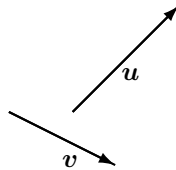
Consider the case where $p = 3$ for now. If you draw an arrow from the point in three dimensional space determined by $(0,0,0)$ to the point (a,b,c) with its tail sitting at the point $(0,0,0)$ and its point at the point (a,b,c) , this arrow is called the **position vector** of the point determined by $u \equiv (a,b,c)$. One way to get to this point is to start at $(0,0,0)$ and move in the direction of the x_1 axis to $(a,0,0)$ and then in the direction of the x_2 axis to $(a,b,0)$ and finally in the direction of the x_3 axis to (a,b,c) . It is evident that the same arrow (vector) would result if you began at the point $v \equiv (d,e,f)$, moved in the direction of the x_1 axis to $(d+a,e,f)$, then in the direction of the x_2 axis to $(d+a,e+b,f)$, and finally in the x_3 direction to $(d+a,e+b,f+c)$ only this time, the arrow would have its tail sitting at the point determined by $v \equiv (d,e,f)$ and its point at $(d+a,e+b,f+c)$. It is said to be the same arrow (vector) because it will point in the same direction and have the same length. It is like you took an actual arrow, the sort of thing you shoot with a bow, and moved it from one location to another keeping it pointing the same direction. This is illustrated in the following picture in which $v + u$ is illustrated. Note the parallelogram determined in the picture by the vectors u and v .



Thus the geometric significance of $(d,e,f) + (a,b,c) = (d+a,e+b,f+c)$ is this. You start with the position vector of the point (d,e,f) and at its point, you place the vector determined by (a,b,c) with its tail at (d,e,f) . Then the point of this last vector will be $(d+a,e+b,f+c)$. This is the geometric significance of vector addition. Also, as shown in the picture, $u + v$ is the directed diagonal of the parallelogram determined by the two vectors u and v .

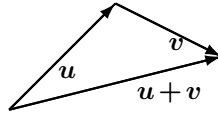
The following example is art.

Example 13.3.1 Here is a picture of two vectors u and v .

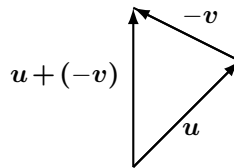


Sketch a picture of $u + v$, $u - v$, and $u + 2v$.

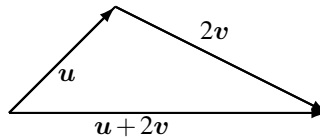
First here is a picture of $u + v$. You first draw u and then at the point of u you place the tail of v as shown. Then $u + v$ is the vector which results which is drawn in the following pretty picture.



Next consider $u - v$. This means $u + (-v)$. From the above geometric description of vector addition, $-v$ is the vector which has the same length but which points in the opposite direction to v . Here is a picture.



Finally consider the vector $u + 2v$. Here is a picture of this one also.



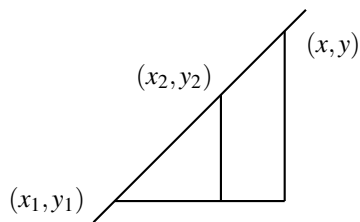
One can think of the point in \mathbb{R}^p identified as (x_1, \dots, x_p) . One can also write this as $(x_1 \ \cdots \ x_p)$. Usually we have in mind a point when there are commas and a vector when it is written as a $1 \times p$ matrix as just illustrated, but it doesn't matter much because the list of numbers with a comma just represents a vector extending from $\mathbf{0}$ to the given point. Therefore, I will use either notation interchangeably.

13.4 Lines

To begin with consider the case $p = 1, 2$. In the case where $p = 1$, the only line is just $\mathbb{R}^1 = \mathbb{R}$. Therefore, if x_1 and x_2 are two different points in \mathbb{R} , consider

$$x = x_1 + t(x_2 - x_1)$$

where $t \in \mathbb{R}$ and the totality of all such points will give \mathbb{R} . You see that you can always solve the above equation for t , showing that every point on \mathbb{R} is of this form. Now consider the plane. Does a similar formula hold? Let (x_1, y_1) and (x_2, y_2) be two different points in \mathbb{R}^2 which are contained in a line l . Suppose that $x_1 \neq x_2$. Then if (x, y) is an arbitrary point on l ,



Now by similar triangles,

$$m \equiv \frac{y_2 - y_1}{x_2 - x_1} = \frac{y - y_1}{x - x_1}$$

and so the point slope form of the line, l , is given as

$$y - y_1 = m(x - x_1).$$

If t is defined by

$$x = x_1 + t(x_2 - x_1),$$

you obtain this equation along with

$$y = y_1 + mt(x_2 - x_1) = y_1 + t(y_2 - y_1).$$

Therefore,

$$(x, y) = (x_1, y_1) + t(x_2 - x_1, y_2 - y_1).$$

If $x_1 = x_2$, then in place of the point slope form above, $x = x_1$. Since the two given points are different, $y_1 \neq y_2$ and so you still obtain the above formula for the line. Because of this, the following is the definition of a line in \mathbb{R}^p .

Definition 13.4.1 A line in \mathbb{R}^p containing the two different points \mathbf{x}^1 and \mathbf{x}^2 is the collection of points of the form

$$\mathbf{x} = \mathbf{x}^1 + t(\mathbf{x}^2 - \mathbf{x}^1)$$

where $t \in \mathbb{R}$. This is known as a **parametric equation** and the variable t is called the **parameter**.

Often t denotes time in applications to Physics. Note this definition agrees with the usual notion of a line in two dimensions and so this is consistent with earlier concepts.

Lemma 13.4.2 Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$ with $\mathbf{a} \neq \mathbf{0}$. Then $\mathbf{x} = t\mathbf{a} + \mathbf{b}$, $t \in \mathbb{R}$, is a line.

Proof: Let $\mathbf{x}^1 = \mathbf{b}$ and let $\mathbf{x}^2 - \mathbf{x}^1 = \mathbf{a}$ so that $\mathbf{x}^2 \neq \mathbf{x}^1$. Then $t\mathbf{a} + \mathbf{b} = \mathbf{x}^1 + t(\mathbf{x}^2 - \mathbf{x}^1)$ and so $\mathbf{x} = t\mathbf{a} + \mathbf{b}$ is a line containing the two different points \mathbf{x}^1 and \mathbf{x}^2 . ■

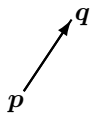
Definition 13.4.3 The vector \mathbf{a} in the above lemma is called a **direction vector** for the line.

Definition 13.4.4 Let \mathbf{p} and \mathbf{q} be two points in \mathbb{R}^p , $\mathbf{p} \neq \mathbf{q}$. The **directed line segment** from \mathbf{p} to \mathbf{q} , denoted by $\overrightarrow{\mathbf{pq}}$, is defined to be the collection of points

$$\mathbf{x} = \mathbf{p} + t(\mathbf{q} - \mathbf{p}), t \in [0, 1]$$

with the direction corresponding to increasing t . In the definition, when $t = 0$, the point \mathbf{p} is obtained and as t increases other points on this line segment are obtained until when $t = 1$, you get the point \mathbf{q} . This is what is meant by saying the direction corresponds to increasing t .

Think of $\overrightarrow{\mathbf{pq}}$ as an arrow whose point is on \mathbf{q} and whose base is at \mathbf{p} as shown in the following picture.



This line segment is a part of a line from the above Definition.

Example 13.4.5 Find a parametric equation for the line through the points $(1, 2, 0)$ and $(2, -4, 6)$.

Use the definition of a line given above to write

$$(x, y, z) = (1, 2, 0) + t(1, -6, 6), t \in \mathbb{R}.$$

The vector $(1, -6, 6)$ is obtained by $(2, -4, 6) - (1, 2, 0)$ as indicated above.

The reason for the word, “a”, rather than the word, “the” is there are infinitely many different parametric equations for the same line. To see this replace t with $3s$. Then you obtain a parametric equation for the same line because the same set of points is obtained. The difference is they are obtained from different values of the parameter. What happens is this: The line is a set of points but the parametric description gives more information than that. It tells how the points are obtained. Obviously, there are many ways to trace out a given set of points and each of these ways corresponds to a different parametric equation for the line.

Example 13.4.6 Find a parametric equation for the line which contains the point $(1, 2, 0)$ and has direction vector $(1, 2, 1)$.

From the above this is just

$$(x, y, z) = (1, 2, 0) + t(1, 2, 1), t \in \mathbb{R}. \quad (13.11)$$

Sometimes people elect to write a line like the above in the form

$$x = 1 + t, y = 2 + 2t, z = t, t \in \mathbb{R}. \quad (13.12)$$

This is a set of scalar parametric equations which amounts to the same thing as 13.11.

There is one other form for a line which is sometimes considered useful. It is the so called symmetric form. Consider the line of 13.12. You can solve for the parameter t to write

$$t = x - 1, t = \frac{y - 2}{2}, t = z.$$

Therefore,

$$x - 1 = \frac{y - 2}{2} = z.$$

This is the symmetric form of the line.

Example 13.4.7 Suppose the *symmetric form of a line* is

$$\frac{x-2}{3} = \frac{y-1}{2} = z+3.$$

Find the line in parametric form.

Let $t = \frac{x-2}{3}, t = \frac{y-1}{2}$ and $t = z+3$. Then solving for x, y, z , you get

$$x = 3t + 2, y = 2t + 1, z = t - 3, t \in \mathbb{R}.$$

Written in terms of vectors this is

$$(2, 1, -3) + t(3, 2, 1) = (x, y, z), t \in \mathbb{R}.$$

I don't understand why anyone would care about the symmetric form of a line if a parametric description is available. Indeed, in linear algebra, you do row operations to express the solution not as a symmetric equation but parametrically.

13.5 Distance in \mathbb{R}^p

How is distance between two points in \mathbb{R}^p defined?

Definition 13.5.1 Let $\mathbf{x} = (x_1, \dots, x_p)$ and $\mathbf{y} = (y_1, \dots, y_p)$ be two points in \mathbb{R}^p . Then $|\mathbf{x} - \mathbf{y}|$ indicates the distance between these points and is defined as

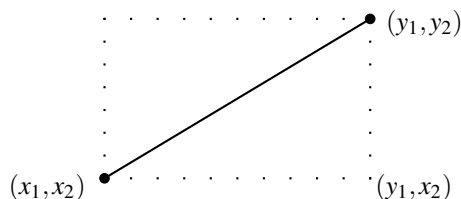
$$\text{distance between } \mathbf{x} \text{ and } \mathbf{y} \equiv |\mathbf{x} - \mathbf{y}| \equiv \left(\sum_{k=1}^p |x_k - y_k|^2 \right)^{1/2}.$$

This is called the **distance formula**. Thus $|\mathbf{x}| \equiv |\mathbf{x} - \mathbf{0}|$. The symbol $B(\mathbf{a}, r)$ is defined by

$$B(\mathbf{a}, r) \equiv \{\mathbf{x} \in \mathbb{R}^p : |\mathbf{x} - \mathbf{a}| < r\}.$$

This is called an **open ball** of radius r centered at \mathbf{a} . It gives all the points in \mathbb{R}^p which are closer to \mathbf{a} than r .

First of all note this is a generalization of the notion of distance in \mathbb{R} . There the distance between two points x and y was given by the absolute value of their difference. Thus $|x - y|$ is equal to the distance between these two points on \mathbb{R} . Now $|x - y| = \left((x - y)^2 \right)^{1/2}$ where the square root is always the positive square root. Thus it is the same formula as the above definition except there is only one term in the sum. Geometrically, this is the right way to define distance which is seen from the Pythagorean theorem. Consider the following picture in the case that $p = 2$.

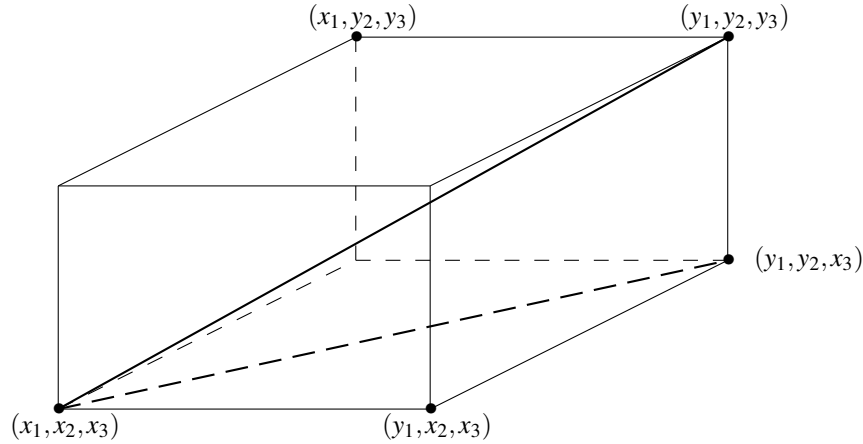


There are two points in the plane whose Cartesian coordinates are (x_1, x_2) and (y_1, y_2) respectively. Then the solid line joining these two points is the hypotenuse of a right triangle which is half of the rectangle shown in dotted lines. What is its length? Note the lengths of the sides of this triangle are $|y_1 - x_1|$ and $|y_2 - x_2|$. Therefore, the Pythagorean theorem implies the length of the hypotenuse equals

$$\left(|y_1 - x_1|^2 + |y_2 - x_2|^2\right)^{1/2} = \left((y_1 - x_1)^2 + (y_2 - x_2)^2\right)^{1/2}$$

which is just the formula for the distance given above.

Now suppose $p = 3$ and let (x_1, x_2, x_3) and (y_1, y_2, y_3) be two points in \mathbb{R}^3 . Consider the following picture in which one of the solid lines joins the two points and a dashed line joins the points (x_1, x_2, x_3) and (y_1, y_2, x_3) .



By the Pythagorean theorem, the length of the dashed line joining (x_1, x_2, x_3) and (y_1, y_2, x_3) equals

$$\left((y_1 - x_1)^2 + (y_2 - x_2)^2\right)^{1/2}$$

while the length of the line joining (y_1, y_2, x_3) to (y_1, y_2, y_3) is just $|y_3 - x_3|$. Therefore, by the Pythagorean theorem again, the length of the line joining the points (x_1, x_2, x_3) and (y_1, y_2, y_3) equals

$$\begin{aligned} & \left\{ \left[\left((y_1 - x_1)^2 + (y_2 - x_2)^2 \right)^{1/2} \right]^2 + (y_3 - x_3)^2 \right\}^{1/2} \\ &= \left((y_1 - x_1)^2 + (y_2 - x_2)^2 + (y_3 - x_3)^2 \right)^{1/2}, \end{aligned}$$

which is again just the distance formula above.

This completes the argument that the above definition is reasonable. Of course you cannot continue drawing pictures in ever higher dimensions but there is no problem with the formula for distance in any number of dimensions. Here is an example.

Example 13.5.2 Find the distance between the points in \mathbb{R}^4 ,

$$\mathbf{a} = (1, 2, -4, 6), \mathbf{b} = (2, 3, -1, 0)$$

Use the distance formula and write

$$|\mathbf{a} - \mathbf{b}|^2 = (1 - 2)^2 + (2 - 3)^2 + (-4 - (-1))^2 + (6 - 0)^2 = 47$$

Therefore, $|\mathbf{a} - \mathbf{b}| = \sqrt{47}$.

All this amounts to defining the distance between two points as the length of a straight line joining these two points. However, there is nothing sacred about using straight lines. One could define the distance to be the length of some other sort of line joining these points. It won't be done very much in this book.

Another convention which is usually followed, especially in \mathbb{R}^2 and \mathbb{R}^3 is to denote the first component of a point in \mathbb{R}^2 by x and the second component by y . In \mathbb{R}^3 it is customary to denote the first and second components as just described while the third component is called z .

Example 13.5.3 Describe the points which are at the same distance between $(1, 2, 3)$ and $(0, 1, 2)$.

Let (x, y, z) be such a point. Then

$$\sqrt{(x-1)^2 + (y-2)^2 + (z-3)^2} = \sqrt{x^2 + (y-1)^2 + (z-2)^2}.$$

Squaring both sides

$$(x-1)^2 + (y-2)^2 + (z-3)^2 = x^2 + (y-1)^2 + (z-2)^2$$

and so

$$x^2 - 2x + 14 + y^2 - 4y + z^2 - 6z = x^2 + y^2 - 2y + 5 + z^2 - 4z$$

which implies

$$-2x + 14 - 4y - 6z = -2y + 5 - 4z$$

hence

$$2x + 2y + 2z = -9. \quad (13.13)$$

Since these steps are reversible, the set of points which is at the same distance from the two given points consists of the points (x, y, z) such that 13.13 holds.

The following lemma is fundamental. It is a form of the Cauchy Schwarz inequality.

Lemma 13.5.4 Let $\mathbf{x} = (x_1, \dots, x_p)$ and $\mathbf{y} = (y_1, \dots, y_p)$ be two points in \mathbb{R}^p . Then

$$\left| \sum_{i=1}^p x_i y_i \right| \leq |\mathbf{x}| |\mathbf{y}| = \left(\sum_{i=1}^p |x_i|^2 \right)^{1/2} \left(\sum_{i=1}^p |y_i|^2 \right)^{1/2}. \quad (13.14)$$

Proof: Let θ be either 1 or -1 such that

$$\theta \sum_{i=1}^p x_i y_i = \sum_{i=1}^p x_i (\theta y_i) = \left| \sum_{i=1}^p x_i y_i \right|$$

and consider $p(t) \equiv \sum_{i=1}^p (x_i + t\theta y_i)^2$. Then for all $t \in \mathbb{R}$,

$$0 \leq p(t) = \sum_{i=1}^p x_i^2 + 2t \sum_{i=1}^p x_i \theta y_i + t^2 \sum_{i=1}^p y_i^2 = |\mathbf{x}|^2 + 2t \sum_{i=1}^p x_i \theta y_i + t^2 |\mathbf{y}|^2$$

If $|\mathbf{y}| = 0$ then 13.14 is obviously true because both sides equal zero. Therefore, assume $|\mathbf{y}| \neq 0$ and then $p(t)$ is a polynomial of degree two whose graph opens up. Therefore, it either has no zeroes, two zeros or one repeated zero. If it has two zeros, the above inequality must be violated because in this case the graph must dip below the x axis. Therefore, it either has no zeros or exactly one. From the quadratic formula this happens exactly when

$$4 \left(\sum_{i=1}^p x_i \theta y_i \right)^2 - 4 |\mathbf{x}|^2 |\mathbf{y}|^2 \leq 0$$

and so

$$\theta \sum_{i=1}^p x_i y_i = \left| \sum_{i=1}^p x_i y_i \right| \leq |\mathbf{x}| |\mathbf{y}|$$

as claimed. This proves the inequality. ■

There are certain properties of the distance which are obvious. Two of them which follow directly from the definition are

$$|\mathbf{x} - \mathbf{y}| = |\mathbf{y} - \mathbf{x}|,$$

$$|\mathbf{x} - \mathbf{y}| \geq 0 \text{ and equals } 0 \text{ only if } \mathbf{y} = \mathbf{x}.$$

The third fundamental property of distance is known as the triangle inequality. Recall that in any triangle the sum of the lengths of two sides is always at least as large as the third side. The following corollary is equivalent to this simple statement.

Corollary 13.5.5 *Let \mathbf{x}, \mathbf{y} be points of \mathbb{R}^p . Then*

$$|\mathbf{x} + \mathbf{y}| \leq |\mathbf{x}| + |\mathbf{y}|.$$

Proof: Using the Cauchy Schwarz inequality, Lemma 13.5.4,

$$\begin{aligned} |\mathbf{x} + \mathbf{y}|^2 &\equiv \sum_{i=1}^p (x_i + y_i)^2 = \sum_{i=1}^p x_i^2 + 2 \sum_{i=1}^p x_i y_i + \sum_{i=1}^p y_i^2 \\ &\leq |\mathbf{x}|^2 + 2 |\mathbf{x}| |\mathbf{y}| + |\mathbf{y}|^2 = (|\mathbf{x}| + |\mathbf{y}|)^2 \end{aligned}$$

and so upon taking square roots of both sides,

$$|\mathbf{x} + \mathbf{y}| \leq |\mathbf{x}| + |\mathbf{y}| \quad \blacksquare$$

13.6 Geometric Meaning of Scalar Multiplication in \mathbb{R}^3

As discussed earlier, $\mathbf{x} = (x_1, x_2, x_3)$ determines a vector. You draw the line from $\mathbf{0}$ to \mathbf{x} placing the point of the vector on \mathbf{x} . What is the length of this vector? The length of this vector is defined to equal $|\mathbf{x}|$ as in Definition 13.5.1. Thus the length of \mathbf{x} equals $\sqrt{x_1^2 + x_2^2 + x_3^2}$. When you multiply \mathbf{x} by a scalar α , you get $(\alpha x_1, \alpha x_2, \alpha x_3)$ and the length of this vector is defined as

$$\sqrt{(\alpha x_1)^2 + (\alpha x_2)^2 + (\alpha x_3)^2} = |\alpha| \sqrt{x_1^2 + x_2^2 + x_3^2}.$$

Thus the following holds.

$$|\alpha x| = |\alpha| |x|.$$

In other words, multiplication by a scalar magnifies the length of the vector. What about the direction? You should convince yourself by drawing a picture that if α is negative, it causes the resulting vector to point in the opposite direction while if $\alpha > 0$ it preserves the direction the vector points. One way to see this is to first observe that if $\alpha \neq 1$, then x and αx are both points on the same line going through $\mathbf{0}$. Note that there is no change in this when you replace \mathbb{R}^3 with \mathbb{R}^p .

13.7 Exercises

1. Verify all the properties 13.3-13.10.

2. Compute the following

$$(a) \begin{pmatrix} 1 & 2 & 3 & -2 \end{pmatrix} + 6 \begin{pmatrix} 2 & 1 & -2 & 7 \end{pmatrix}$$

$$(b) -2 \begin{pmatrix} 1 & 2 & -2 \end{pmatrix} + 6 \begin{pmatrix} 2 & 1 & -2 \end{pmatrix}$$

3. Find symmetric equations for the line through the points $(2, 2, 4)$ and $(-2, 3, 1)$. Dumb idea but do it anyway.
4. Find symmetric equations for the line through the points $(1, 2, 4)$ and $(-2, 1, 1)$. Dumb idea but do it anyway.
5. Symmetric equations for a line are given. Find parametric equations of the line. This goes the right direction.

$$(a) \frac{x+1}{3} = \frac{2y+3}{2} = z+7$$

$$(b) \frac{2x-1}{3} = \frac{2y+3}{6} = z-7$$

6. The first point given is a point contained in the line. The second point given is a direction vector for the line. Find parametric equations for the line, determined by this information.

$$(a) (1, 2, 1), (2, 0, 3)$$

$$(b) (1, 0, 1), (1, 1, 3)$$

$$(c) (1, 2, 0), (1, 1, 0)$$

7. Parametric equations for a line are given. Determine a direction vector for this line.

$$(a) x = 1 + 2t, y = 3 - t, z = 5 + 3t$$

$$(b) x = 1 + t, y = 3 + 3t, z = 5 - t$$

8. A line contains the given two points. Find parametric equations for this line. Identify the direction vector.

$$(a) (0, 1, 0), (2, 1, 2)$$

$$(b) (0, 1, 1), (2, 5, 0)$$

9. Describe in words how to get to the points described by the ordered pairs.
 - (a) $(1, 2)$
 - (b) $(-2, -2)$
10. Does it make sense to write $\begin{pmatrix} 1 & 2 \end{pmatrix} + \begin{pmatrix} 2 & 3 & 1 \end{pmatrix}$? Explain.
11. Describe in words how to get to the point in \mathbb{R}^3 denoted by the ordered triples.
 - (a) $(1, 2, 0)$
 - (b) $(-2, -2, 1)$
 - (c) $(-2, 3, -2)$
12. You are given two points in \mathbb{R}^3 , $(4, 5, -4)$ and $(2, 3, 0)$. Show the distance from the point $(3, 4, -2)$ to the first of these points is the same as the distance from this point to the second of the original pair of points. Note that $3 = \frac{4+2}{2}$, $4 = \frac{5+3}{2}$. Obtain a theorem which will be valid for general pairs of points (x, y, z) and (x_1, y_1, z_1) and prove your theorem using the distance formula.
13. A sphere is the set of all points which are at a given distance from a single given point. Find an equation for the sphere which is the set of all points that are at a distance of 4 from the point $(1, 2, 3)$ in \mathbb{R}^3 .
14. A parabola is the set of all points (x, y) in the plane such that the distance from the point (x, y) to a given point (x_0, y_0) equals the distance from (x, y) to a given line. The point (x_0, y_0) is called the **focus** and the line is called the **directrix**. Find the equation of the parabola which results from the line $y = l$ and (x_0, y_0) a given focus with $y_0 < l$. Repeat for $y_0 > l$.
15. Suppose the distance between (x, y) and (x', y') were defined to equal the larger of the two numbers $|x - x'|$ and $|y - y'|$. Draw a picture of the sphere centered at the point $(0, 0)$ if this notion of distance is used.
16. Repeat the same problem except this time let the distance between the two points be $|x - x'| + |y - y'|$.
17. If (x_1, y_1, z_1) and (x_2, y_2, z_2) are two points such that $|(x_i, y_i, z_i)| = 1$ for $i = 1, 2$, show that in terms of the usual distance, $\left| \left(\frac{x_1+x_2}{2}, \frac{y_1+y_2}{2}, \frac{z_1+z_2}{2} \right) \right| < 1$. What would happen if you used the way of measuring distance given in Problem 15 ($|(x, y, z)| = \text{maximum of } |z|, |x|, |y|$)?
18. Give a simple description using the distance formula of the set of points which are at an equal distance between the two points (x_1, y_1, z_1) and (x_2, y_2, z_2) .
19. Suppose you are given two points $(-a, 0)$ and $(a, 0)$ in \mathbb{R}^2 and a number $r > 2a$. The set of points described by

$$\{(x, y) \in \mathbb{R}^2 : |(x, y) - (-a, 0)| + |(x, y) - (a, 0)| = r\}$$

is known as an ellipse. The two given points are known as the **focus points** of the ellipse. Find α and β such that this is in the form $\left(\frac{x}{\alpha}\right)^2 + \left(\frac{y}{\beta}\right)^2 = 1$. This is a nice exercise in messy algebra.

20. Suppose you are given two points $(-a, 0)$ and $(a, 0)$ in \mathbb{R}^2 and a number $r < 2a$. The set of points described by

$$\{(x, y) \in \mathbb{R}^2 : |(x, y) - (-a, 0)| - |(x, y) - (a, 0)| = r\}$$

is known as **hyperbola**. The two given points are known as the **focus points** of the hyperbola. Simplify this to the form $\left(\frac{x}{a}\right)^2 - \left(\frac{y}{b}\right)^2 = 1$. This is a nice exercise in messy algebra.

21. Let (x_1, y_1) and (x_2, y_2) be two points in \mathbb{R}^2 . Give a simple description using the distance formula of the perpendicular bisector of the line segment joining these two points. Thus you want all points (x, y) such that $|(x, y) - (x_1, y_1)| = |(x, y) - (x_2, y_2)|$.
22. Show that $|\alpha x| = |\alpha| |x|$ whenever $x \in \mathbb{R}^p$ for any positive integer p .

Chapter 14

Vector Products

14.1 The Dot Product

There are two ways of multiplying vectors which are of great importance in applications. The first of these is called the **dot product**, also called the **scalar product** and sometimes the **inner product**.

Definition 14.1.1 Let \mathbf{a}, \mathbf{b} be two vectors in \mathbb{R}^p define $\mathbf{a} \cdot \mathbf{b}$ as

$$\mathbf{a} \cdot \mathbf{b} \equiv \sum_{k=1}^p a_k b_k.$$

With this definition, there are several important properties satisfied by the dot product. In the statement of these properties, α and β will denote scalars and $\mathbf{a}, \mathbf{b}, \mathbf{c}$ will denote vectors.

Proposition 14.1.2 The dot product satisfies the following properties.

$$\mathbf{a} \cdot \mathbf{b} = \mathbf{b} \cdot \mathbf{a} \quad (14.1)$$

$$\mathbf{a} \cdot \mathbf{a} \geq 0 \text{ and equals zero if and only if } \mathbf{a} = \mathbf{0} \quad (14.2)$$

$$(\alpha \mathbf{a} + \beta \mathbf{b}) \cdot \mathbf{c} = \alpha (\mathbf{a} \cdot \mathbf{c}) + \beta (\mathbf{b} \cdot \mathbf{c}) \quad (14.3)$$

$$\mathbf{c} \cdot (\alpha \mathbf{a} + \beta \mathbf{b}) = \alpha (\mathbf{c} \cdot \mathbf{a}) + \beta (\mathbf{c} \cdot \mathbf{b}) \quad (14.4)$$

$$|\mathbf{a}|^2 = \mathbf{a} \cdot \mathbf{a} \quad (14.5)$$

You should verify these properties. Also be sure you understand that 14.4 follows from the first three and is therefore redundant. It is listed here for the sake of convenience.

Example 14.1.3 Find $(1, 2, 0, -1) \cdot (0, 1, 2, 3)$.

This equals $0 + 2 + 0 + -3 = -1$.

Example 14.1.4 Find the magnitude of $\mathbf{a} = (2, 1, 4, 2)$. That is, find $|\mathbf{a}|$.

This is $\sqrt{(2, 1, 4, 2) \cdot (2, 1, 4, 2)} = 5$.

The dot product satisfies the **CauchySchwarz inequality**. It has already been proved but here is another proof. This proof will be based only on the above axioms for the dot product.

Theorem 14.1.5 *The dot product satisfies the inequality*

$$|\mathbf{a} \cdot \mathbf{b}| \leq |\mathbf{a}| |\mathbf{b}|. \quad (14.6)$$

Furthermore equality is obtained if and only if one of \mathbf{a} or \mathbf{b} is a scalar multiple of the other.

Proof: First note that if $\mathbf{b} = \mathbf{0}$, both sides of 14.6 equal zero and so the inequality holds in this case. Indeed,

$$\mathbf{a} \cdot \mathbf{0} = \mathbf{a} \cdot (\mathbf{0} + \mathbf{0}) = \mathbf{a} \cdot \mathbf{0} + \mathbf{a} \cdot \mathbf{0}$$

so $\mathbf{a} \cdot \mathbf{0} = 0$. Therefore, it will be assumed in what follows that $\mathbf{b} \neq \mathbf{0}$.

Define a function of $t \in \mathbb{R}$

$$f(t) = (\mathbf{a} + t\mathbf{b}) \cdot (\mathbf{a} + t\mathbf{b}).$$

Then by 14.2, $f(t) \geq 0$ for all $t \in \mathbb{R}$. Also from 14.3, 14.4, 14.1, and 14.5

$$\begin{aligned} f(t) &= \mathbf{a} \cdot (\mathbf{a} + t\mathbf{b}) + t\mathbf{b} \cdot (\mathbf{a} + t\mathbf{b}) = \mathbf{a} \cdot \mathbf{a} + t(\mathbf{a} \cdot \mathbf{b}) + t\mathbf{b} \cdot \mathbf{a} + t^2\mathbf{b} \cdot \mathbf{b} \\ &= |\mathbf{a}|^2 + 2t(\mathbf{a} \cdot \mathbf{b}) + |\mathbf{b}|^2 t^2. \end{aligned}$$

Then solve $f'(t) = 0$ for t . This gives $t = \frac{-(\mathbf{a} \cdot \mathbf{b})}{|\mathbf{b}|^2}$. Plug this value of t into the formula for $f(t)$. Then

$$\begin{aligned} 0 &\leq |\mathbf{a}|^2 + 2 \left(\frac{-(\mathbf{a} \cdot \mathbf{b})}{|\mathbf{b}|^2} \right) (\mathbf{a} \cdot \mathbf{b}) + |\mathbf{b}|^2 \left(\frac{-(\mathbf{a} \cdot \mathbf{b})}{|\mathbf{b}|^2} \right)^2 \\ &= |\mathbf{a}|^2 - \frac{2(\mathbf{a} \cdot \mathbf{b})^2}{|\mathbf{b}|^2} + \frac{(\mathbf{a} \cdot \mathbf{b})^2}{|\mathbf{b}|^2} = |\mathbf{a}|^2 - \frac{(\mathbf{a} \cdot \mathbf{b})^2}{|\mathbf{b}|^2} = f \left(\frac{-(\mathbf{a} \cdot \mathbf{b})}{|\mathbf{b}|^2} \right) \end{aligned} \quad (14.7)$$

which shows

$$(\mathbf{a} \cdot \mathbf{b})^2 \leq |\mathbf{a}|^2 |\mathbf{b}|^2, \quad |(\mathbf{a} \cdot \mathbf{b})| \leq |\mathbf{a}| |\mathbf{b}|.$$

From properties of the dot product, equality holds in 14.6 whenever one of the vectors is a scalar multiple of the other. It only remains to verify this is the only way equality can occur. If either vector equals zero, then one is a multiple of the other. If equality holds, in the inequality, then $f(t) = 0$ from 14.7. Therefore, for t the point where minimum of f is achieved, $(\mathbf{a} + t\mathbf{b}) \cdot (\mathbf{a} + t\mathbf{b}) = 0$ and so $\mathbf{a} = -t\mathbf{b}$. ■

You should note that the entire argument was based only on the properties of the dot product listed in 14.1 - 14.5. This means that whenever something satisfies these axioms, the Cauchy Schwartz inequality holds. There are many other instances of these properties besides vectors in \mathbb{R}^p .

The Cauchy Schwartz inequality allows a proof of the **triangle inequality** for distances in \mathbb{R}^p in much the same way as the triangle inequality for the absolute value.

Theorem 14.1.6 (Triangle inequality) For $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$

$$|\mathbf{a} + \mathbf{b}| \leq |\mathbf{a}| + |\mathbf{b}| \quad (14.8)$$

and equality holds if and only if one of the vectors is a nonnegative scalar multiple of the other. Also

$$||\mathbf{a}| - |\mathbf{b}|| \leq |\mathbf{a} - \mathbf{b}| \quad (14.9)$$

Proof: By properties of the dot product and the Cauchy Schwarz inequality,

$$\begin{aligned} |\mathbf{a} + \mathbf{b}|^2 &= (\mathbf{a} + \mathbf{b}) \cdot (\mathbf{a} + \mathbf{b}) = (\mathbf{a} \cdot \mathbf{a}) + (\mathbf{a} \cdot \mathbf{b}) + (\mathbf{b} \cdot \mathbf{a}) + (\mathbf{b} \cdot \mathbf{b}) \\ &= |\mathbf{a}|^2 + 2(\mathbf{a} \cdot \mathbf{b}) + |\mathbf{b}|^2 \leq |\mathbf{a}|^2 + 2|\mathbf{a} \cdot \mathbf{b}| + |\mathbf{b}|^2 \\ &\leq |\mathbf{a}|^2 + 2|\mathbf{a}||\mathbf{b}| + |\mathbf{b}|^2 = (|\mathbf{a}| + |\mathbf{b}|)^2. \end{aligned}$$

Taking square roots of both sides you obtain 14.8.

It remains to consider when equality occurs. If either vector equals zero, then that vector equals zero times the other vector and the claim about when equality occurs is verified. Therefore, it can be assumed both vectors are nonzero. To get equality in the second inequality above, Theorem 14.1.5 implies one of the vectors must be a multiple of the other. Say $\mathbf{b} = \alpha \mathbf{a}$. If $\alpha < 0$ then equality cannot occur in the first inequality because in this case

$$(\mathbf{a} \cdot \mathbf{b}) = \alpha |\mathbf{a}|^2 < 0 < |\alpha| |\mathbf{a}|^2 = |\mathbf{a} \cdot \mathbf{b}|$$

Therefore, $\alpha \geq 0$.

To get the other form of the triangle inequality, $\mathbf{a} = \mathbf{a} - \mathbf{b} + \mathbf{b}$ so

$$|\mathbf{a}| = |\mathbf{a} - \mathbf{b} + \mathbf{b}| \leq |\mathbf{a} - \mathbf{b}| + |\mathbf{b}|.$$

Therefore,

$$|\mathbf{a}| - |\mathbf{b}| \leq |\mathbf{a} - \mathbf{b}| \quad (14.10)$$

Similarly,

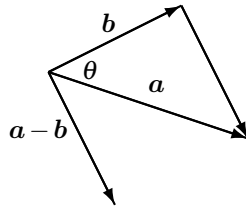
$$|\mathbf{b}| - |\mathbf{a}| \leq |\mathbf{b} - \mathbf{a}| = |\mathbf{a} - \mathbf{b}|. \quad (14.11)$$

It follows from 14.10 and 14.11 that 14.9 holds. This is because $||\mathbf{a}| - |\mathbf{b}||$ equals the left side of either 14.10 or 14.11 and either way, $||\mathbf{a}| - |\mathbf{b}|| \leq |\mathbf{a} - \mathbf{b}|$. ■

14.2 Geometric Significance of the Dot Product

14.2.1 The Angle Between Two Vectors

Given two vectors \mathbf{a} and \mathbf{b} , the included angle is the angle between these two vectors which is less than or equal to 180 degrees. The dot product can be used to determine the included angle between two vectors. To see how to do this, consider the following picture.



By the law of cosines,

$$|\mathbf{a} - \mathbf{b}|^2 = |\mathbf{a}|^2 + |\mathbf{b}|^2 - 2|\mathbf{a}||\mathbf{b}|\cos\theta.$$

Also from the properties of the dot product,

$$|\mathbf{a} - \mathbf{b}|^2 = (\mathbf{a} - \mathbf{b}) \cdot (\mathbf{a} - \mathbf{b}) = |\mathbf{a}|^2 + |\mathbf{b}|^2 - 2\mathbf{a} \cdot \mathbf{b}$$

and so comparing the above two formulas,

$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}||\mathbf{b}|\cos\theta. \quad (14.12)$$

In words, the dot product of two vectors equals the product of the magnitude of the two vectors multiplied by the cosine of the included angle. Note this gives a geometric description of the dot product which does not depend explicitly on the coordinates of the vectors.

Example 14.2.1 Find the angle between the vectors $2\mathbf{i} + \mathbf{j} - \mathbf{k}$ and $3\mathbf{i} + 4\mathbf{j} + \mathbf{k}$.

The dot product of these two vectors equals $6 + 4 - 1 = 9$ and the norms are

$$\sqrt{4 + 1 + 1} = \sqrt{6}$$

and $\sqrt{9 + 16 + 1} = \sqrt{26}$. Therefore, from 14.12 the cosine of the included angle equals

$$\cos\theta = \frac{9}{\sqrt{26}\sqrt{6}} = .72058$$

Now the cosine is known, the angle can be determined by solving the equation $\cos\theta = .72058$. This will involve using a calculator or a table of trigonometric functions. The answer is $\theta = .76616$ radians or in terms of degrees, $\theta = .76616 \times \frac{360}{2\pi} = 43.898^\circ$. Recall how this last computation is done. Set up a proportion $\frac{x}{.76616} = \frac{360}{2\pi}$ because 360° corresponds to 2π radians. However, in calculus, you should get used to thinking in terms of radians and not degrees. This is because all the important calculus formulas are defined in terms of radians.

Example 14.2.2 Let \mathbf{u}, \mathbf{v} be two vectors whose magnitudes are equal to 3 and 4 respectively and such that if they are placed in standard position with their tails at the origin, the angle between \mathbf{u} and the positive x axis equals 30° and the angle between \mathbf{v} and the positive x axis is -30° . Find $\mathbf{u} \cdot \mathbf{v}$.

From the geometric description of the dot product in 14.12

$$\mathbf{u} \cdot \mathbf{v} = 3 \times 4 \times \cos(60^\circ) = 3 \times 4 \times 1/2 = 6.$$

Observation 14.2.3 Two vectors are said to be **perpendicular** if the included angle is $\pi/2$ radians (90°). You can tell if two nonzero vectors are perpendicular by simply taking their dot product. If the answer is zero, this means they are perpendicular because $\cos\theta = 0$.

Example 14.2.4 Determine whether the two vectors $2\mathbf{i} + \mathbf{j} - \mathbf{k}$ and $1\mathbf{i} + 3\mathbf{j} + 5\mathbf{k}$ are perpendicular.

When you take this dot product you get $2 + 3 - 5 = 0$ and so these two are indeed perpendicular.

Definition 14.2.5 *When two lines intersect, the angle between the two lines is the smaller of the two angles determined.*

Example 14.2.6 *Find the angle between the two lines, $(1, 2, 0) + t(1, 2, 3)$ and $(0, 4, -3) + t(-1, 2, -3)$.*

These two lines intersect, when $t = 0$ in the first and $t = -1$ in the second. It is only a matter of finding the angle between the direction vectors. One angle determined is given by

$$\cos \theta = \frac{-6}{14} = \frac{-3}{7}. \quad (14.13)$$

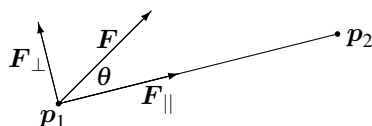
We don't want this angle because it is obtuse. The angle desired is the acute angle given by

$$\cos \theta = \frac{3}{7}.$$

It is obtained by using replacing one of the direction vectors with -1 times it.

14.2.2 Work and Projections

Our first application will be to the concept of work. The physical concept of work does not in any way correspond to the notion of work employed in ordinary conversation. For example, if you were to slide a 150 pound weight off a table which is three feet high and shuffle along the floor for 50 yards, sweating profusely and exerting all your strength to keep the weight from falling on your feet, keeping the height always three feet and then deposit this weight on another three foot high table, the physical concept of work would indicate that the force exerted by your arms did no work during this project even though the muscles in your hands and arms would likely be very tired. The reason for such an unusual definition is that even though your arms exerted considerable force on the weight, enough to keep it from falling, the direction of motion was at right angles to the force they exerted. The only part of a force which does work in the sense of physics is the component of the force in the direction of motion (This is made more precise below.). The work is defined to be the magnitude of the component of this force times the distance over which it acts in the case where this component of force points in the direction of motion and (-1) times the magnitude of this component times the distance in case the force tends to impede the motion. Thus the work done by a force on an object as the object moves from one point to another is a measure of the extent to which the force contributes to the motion. This is illustrated in the following picture in the case where the given force contributes to the motion.



In this picture the force, \mathbf{F} is applied to an object which moves on the straight line from \mathbf{p}_1 to \mathbf{p}_2 . There are two vectors shown, \mathbf{F}_{\parallel} and \mathbf{F}_{\perp} and the picture is intended to indicate that when you add these two vectors you get \mathbf{F} while \mathbf{F}_{\parallel} acts in the direction of motion and \mathbf{F}_{\perp} acts perpendicular to the direction of motion. Only \mathbf{F}_{\parallel} contributes to the work done by \mathbf{F} on the object as it moves from \mathbf{p}_1 to \mathbf{p}_2 . \mathbf{F}_{\parallel} is called the **component of the force** in the direction of motion. From trigonometry, you see the magnitude of \mathbf{F}_{\parallel} should equal $|\mathbf{F}|\cos\theta$. Thus, since \mathbf{F}_{\parallel} points in the direction of the vector from \mathbf{p}_1 to \mathbf{p}_2 , the total work done should equal

$$|\mathbf{F}| |\overrightarrow{p_1 p_2}| \cos \theta = |\mathbf{F}| |\mathbf{p}_2 - \mathbf{p}_1| \cos \theta$$

If the included angle had been obtuse, then the work done by the force, \mathbf{F} on the object would have been negative because in this case, the force tends to impede the motion from \mathbf{p}_1 to \mathbf{p}_2 but in this case, $\cos \theta$ would also be negative and so it is still the case that the work done would be given by the above formula. Thus from the geometric description of the dot product given above, the work equals

$$|\mathbf{F}| |\mathbf{p}_2 - \mathbf{p}_1| \cos \theta = \mathbf{F} \cdot (\mathbf{p}_2 - \mathbf{p}_1).$$

This explains the following definition.

Definition 14.2.7 Let \mathbf{F} be a force acting on an object which moves from the point \mathbf{p}_1 to the point \mathbf{p}_2 . Then the **work** done on the object by the given force equals $\mathbf{F} \cdot (\mathbf{p}_2 - \mathbf{p}_1)$.

The concept of writing a given vector \mathbf{F} in terms of two vectors, one which is parallel to a given vector \mathbf{D} and the other which is perpendicular can also be explained with no reliance on trigonometry, completely in terms of the algebraic properties of the dot product. As before, this is mathematically more significant than any approach involving geometry or trigonometry because it extends to more interesting situations. This is done next.

Theorem 14.2.8 Let \mathbf{F} and \mathbf{D} be nonzero vectors. Then there exist unique vectors \mathbf{F}_{\parallel} and \mathbf{F}_{\perp} such that

$$\mathbf{F} = \mathbf{F}_{\parallel} + \mathbf{F}_{\perp} \tag{14.14}$$

where \mathbf{F}_{\parallel} is a scalar multiple of \mathbf{D} , also referred to as

$$\text{proj}_{\mathbf{D}}(\mathbf{F}),$$

and $\mathbf{F}_{\perp} \cdot \mathbf{D} = 0$. The vector $\text{proj}_{\mathbf{D}}(\mathbf{F})$ is called the **projection** of \mathbf{F} onto \mathbf{D} .

Proof: Suppose 14.14 and $\mathbf{F}_{\parallel} = \alpha \mathbf{D}$. Taking the dot product of both sides with \mathbf{D} and using $\mathbf{F}_{\perp} \cdot \mathbf{D} = 0$, this yields

$$\mathbf{F} \cdot \mathbf{D} = \alpha |\mathbf{D}|^2$$

which requires $\alpha = \mathbf{F} \cdot \mathbf{D} / |\mathbf{D}|^2$. Thus there can be no more than one vector \mathbf{F}_{\parallel} . It follows \mathbf{F}_{\perp} must equal $\mathbf{F} - \mathbf{F}_{\parallel}$. This verifies there can be no more than one choice for both \mathbf{F}_{\parallel} and \mathbf{F}_{\perp} .

Now let

$$\mathbf{F}_{\parallel} \equiv \frac{\mathbf{F} \cdot \mathbf{D}}{|\mathbf{D}|^2} \mathbf{D}$$

and let

$$\mathbf{F}_\perp = \mathbf{F} - \mathbf{F}_\parallel = \mathbf{F} - \frac{\mathbf{F} \cdot \mathbf{D}}{|\mathbf{D}|^2} \mathbf{D}$$

Then $\mathbf{F}_\parallel = \alpha \mathbf{D}$ where $\alpha = \frac{\mathbf{F} \cdot \mathbf{D}}{|\mathbf{D}|^2}$. It only remains to verify $\mathbf{F}_\perp \cdot \mathbf{D} = 0$. But

$$\mathbf{F}_\perp \cdot \mathbf{D} = \mathbf{F} \cdot \mathbf{D} - \frac{\mathbf{F} \cdot \mathbf{D}}{|\mathbf{D}|^2} \mathbf{D} \cdot \mathbf{D} = \mathbf{F} \cdot \mathbf{D} - \mathbf{F} \cdot \mathbf{D} = 0. \blacksquare$$

Example 14.2.9 Let $\mathbf{F} = 2\mathbf{i} + 7\mathbf{j} - 3\mathbf{k}$ Newtons. Find the work done by this force in moving from the point $(1, 2, 3)$ to the point $(-9, -3, 4)$ along the straight line segment joining these points where distances are measured in meters.

According to the definition, this work is

$$(2\mathbf{i} + 7\mathbf{j} - 3\mathbf{k}) \cdot (-10\mathbf{i} - 5\mathbf{j} + \mathbf{k}) = -20 + (-35) + (-3) = -58 \text{ Newton meters.}$$

Note that if the force had been given in pounds and the distance had been given in feet, the units on the work would have been foot pounds. In general, work has units equal to units of a force times units of a length. Instead of writing Newton meter, people write joule because a joule is by definition a Newton meter. That word is pronounced “jewel” and it is the unit of work in the metric system of units. Also be sure you observe that the work done by the force can be negative as in the above example. In fact, work can be either positive, negative, or zero. You just have to do the computations to find out.

Example 14.2.10 Find $\text{proj}_u(\mathbf{v})$ if $\mathbf{u} = 2\mathbf{i} + 3\mathbf{j} - 4\mathbf{k}$ and $\mathbf{v} = \mathbf{i} - 2\mathbf{j} + \mathbf{k}$.

From the above discussion in Theorem 14.2.8, this is just

$$\begin{aligned} & \frac{1}{4 + 9 + 16} (\mathbf{i} - 2\mathbf{j} + \mathbf{k}) \cdot (2\mathbf{i} + 3\mathbf{j} - 4\mathbf{k}) (2\mathbf{i} + 3\mathbf{j} - 4\mathbf{k}) \\ &= \frac{-8}{29} (2\mathbf{i} + 3\mathbf{j} - 4\mathbf{k}) = -\frac{16}{29}\mathbf{i} - \frac{24}{29}\mathbf{j} + \frac{32}{29}\mathbf{k}. \end{aligned}$$

Example 14.2.11 Suppose \mathbf{a} , and \mathbf{b} are vectors and $\mathbf{b}_\perp = \mathbf{b} - \text{proj}_a(\mathbf{b})$. What is the magnitude of \mathbf{b}_\perp in terms of the included angle?

$$\begin{aligned} |\mathbf{b}_\perp|^2 &= (\mathbf{b} - \text{proj}_a(\mathbf{b})) \cdot (\mathbf{b} - \text{proj}_a(\mathbf{b})) = \left(\mathbf{b} - \frac{\mathbf{b} \cdot \mathbf{a}}{|\mathbf{a}|^2} \mathbf{a} \right) \cdot \left(\mathbf{b} - \frac{\mathbf{b} \cdot \mathbf{a}}{|\mathbf{a}|^2} \mathbf{a} \right) \\ &= |\mathbf{b}|^2 - 2 \frac{(\mathbf{b} \cdot \mathbf{a})^2}{|\mathbf{a}|^2} + \left(\frac{\mathbf{b} \cdot \mathbf{a}}{|\mathbf{a}|^2} \right)^2 |\mathbf{a}|^2 = |\mathbf{b}|^2 \left(1 - \frac{(\mathbf{b} \cdot \mathbf{a})^2}{|\mathbf{a}|^2 |\mathbf{b}|^2} \right) \\ &= |\mathbf{b}|^2 (1 - \cos^2 \theta) = |\mathbf{b}|^2 \sin^2(\theta) \end{aligned}$$

where θ is the included angle between \mathbf{a} and \mathbf{b} which is less than π radians. Therefore, taking square roots, $|\mathbf{b}_\perp| = |\mathbf{b}| \sin \theta$.

14.3 Exercises

- Find $(1, 2, 3, 4) \cdot (2, 0, 1, 3)$.
- Use formula 14.12 to verify the Cauchy Schwarz inequality and to show that equality occurs if and only if one of the vectors is a scalar multiple of the other.
- For \mathbf{u}, \mathbf{v} vectors in \mathbb{R}^3 , define the product $\mathbf{u} * \mathbf{v} \equiv u_1 v_1 + 2u_2 v_2 + 3u_3 v_3$. Show the axioms for a dot product all hold for this funny product. Prove the following inequality $|\mathbf{u} * \mathbf{v}| \leq (\mathbf{u} * \mathbf{u})^{1/2} (\mathbf{v} * \mathbf{v})^{1/2}$. **Hint:** Do not try to do this with methods from trigonometry.
- Find the angle between the vectors $3\mathbf{i} - \mathbf{j} - \mathbf{k}$ and $\mathbf{i} + 4\mathbf{j} + 2\mathbf{k}$.
- Find $\text{proj}_{\mathbf{u}}(\mathbf{v})$ where $\mathbf{v} = (1, 0, -2)$ and $\mathbf{u} = (1, 2, 3)$.
- Find $\text{proj}_{\mathbf{u}}(\mathbf{v})$ where $\mathbf{v} = (1, 2, -2, 1)$ and $\mathbf{u} = (1, 2, 3, 0)$.
- Does it make sense to speak of $\text{proj}_{\mathbf{0}}(\mathbf{v})$?
- If \mathbf{F} is a force and \mathbf{D} is a vector, show $\text{proj}_{\mathbf{D}}(\mathbf{F}) = (|\mathbf{F}| \cos \theta) \mathbf{u}$ where \mathbf{u} is the unit vector in the direction of \mathbf{D} , $\mathbf{u} = \mathbf{D}/|\mathbf{D}|$ and θ is the included angle between the two vectors \mathbf{F} and \mathbf{D} . $|\mathbf{F}| \cos \theta$ is sometimes called the component of the force \mathbf{F} in the direction, \mathbf{D} .
- A boy drags a sled for 100 feet along the ground by pulling on a rope which is 20 degrees from the horizontal with a force of 40 pounds. How much work does this force do?
- A girl drags a sled for 200 feet along the ground by pulling on a rope which is 30 degrees from the horizontal with a force of 20 pounds. How much work does this force do?
- How much work in Newton meters does it take to slide a crate 20 meters along a loading dock by pulling on it with a 200 Newton force at an angle of 30° from the horizontal?
- An object moves 10 meters in the direction of \mathbf{j} . There are two forces acting on this object $\mathbf{F}_1 = \mathbf{i} + \mathbf{j} + \mathbf{k}$, and $\mathbf{F}_2 = -5\mathbf{i} + 2\mathbf{j} - 6\mathbf{k}$. Find the total work done on the object by the two forces. **Hint:** You can take the work done by the resultant of the two forces or you can add the work done by each force. Why?
- An object moves 10 meters in the direction of $\mathbf{j} + \mathbf{i}$. There are two forces acting on this object $\mathbf{F}_1 = \mathbf{i} + \mathbf{j} + 2\mathbf{k}$, and $\mathbf{F}_2 = 5\mathbf{i} + 2\mathbf{j} - 6\mathbf{k}$. Find the total work done on the object by the two forces. **Hint:** You can take the work done by the resultant of the two forces or you can add the work done by each force. Why?
- If \mathbf{a}, \mathbf{b} , and \mathbf{c} are vectors. Show that $(\mathbf{b} + \mathbf{c})_{\perp} = \mathbf{b}_{\perp} + \mathbf{c}_{\perp}$ where $\mathbf{b}_{\perp} = \mathbf{b} - \text{proj}_{\mathbf{a}}(\mathbf{b})$.
- Show that $(\mathbf{a} \cdot \mathbf{b}) = \frac{1}{4} [|\mathbf{a} + \mathbf{b}|^2 - |\mathbf{a} - \mathbf{b}|^2]$.
- Prove from the axioms of the dot product the parallelogram identity which asserts that $|\mathbf{a} + \mathbf{b}|^2 + |\mathbf{a} - \mathbf{b}|^2 = 2|\mathbf{a}|^2 + 2|\mathbf{b}|^2$.

17. Suppose f, g are two continuous functions defined on $[0, 1]$. Define

$$(f \cdot g) = \int_0^1 f(x) g(x) dx.$$

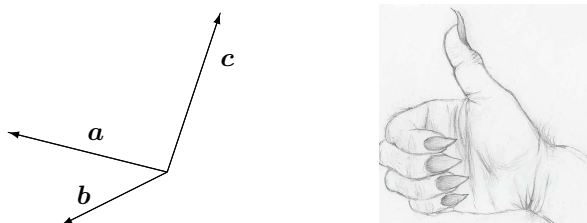
Show this dot product satisfies conditions 14.1 - 14.5. Explain why the Cauchy Schwarz inequality continues to hold in this context and state the Cauchy Schwarz inequality in terms of integrals.

14.4 The Cross Product

The cross product is the other way of multiplying two vectors in \mathbb{R}^3 . It is very different from the dot product in many ways. First the geometric meaning is discussed and then a description in terms of coordinates is given. Both descriptions of the cross product are important. The geometric description is essential in order to understand the applications to physics and geometry while the coordinate description is the only way to practically compute the cross product.

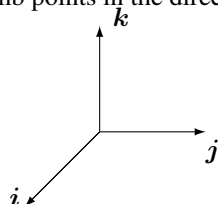
Definition 14.4.1 *Three vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}$ form a right handed system if when you extend the fingers of your right hand along the vector \mathbf{a} and close them in the direction of \mathbf{b} , the thumb points roughly in the direction of \mathbf{c} .*

For an example of a right handed system of vectors, see the following picture.



In this picture the vector \mathbf{c} points upwards from the plane determined by the other two vectors. You should consider how a right hand system would differ from a left hand system. Try using your left hand and you will see that the vector \mathbf{c} would need to point in the opposite direction as it would for a right hand system.

From now on, the vectors $\mathbf{i}, \mathbf{j}, \mathbf{k}$ will always form a right handed system. To repeat, if you extend the fingers of our right hand along \mathbf{i} and close them in the direction \mathbf{j} , the thumb points in the direction of \mathbf{k} .



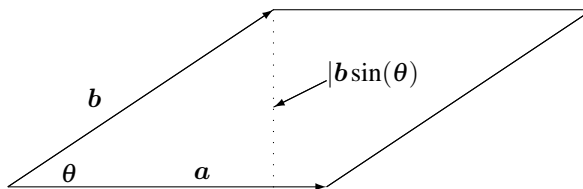
The following is the geometric description of the cross product. It gives both the direction and the magnitude and therefore specifies the vector.

Definition 14.4.2 *Let \mathbf{a} and \mathbf{b} be two vectors in \mathbb{R}^3 . Then $\mathbf{a} \times \mathbf{b}$ is defined by the following two rules.*

1. $|\mathbf{a} \times \mathbf{b}| = |\mathbf{a}| |\mathbf{b}| \sin \theta$ where θ is the included angle.

2. $\mathbf{a} \times \mathbf{b} \cdot \mathbf{a} = 0$, $\mathbf{a} \times \mathbf{b} \cdot \mathbf{b} = 0$, and $\mathbf{a}, \mathbf{b}, \mathbf{a} \times \mathbf{b}$ forms a right hand system.

Note that $|\mathbf{a} \times \mathbf{b}|$ is the area of the parallelogram spanned by \mathbf{a} and \mathbf{b} .



The cross product satisfies the following properties.

$$\mathbf{a} \times \mathbf{b} = -(\mathbf{b} \times \mathbf{a}), \quad \mathbf{a} \times \mathbf{a} = \mathbf{0}, \quad (14.15)$$

For α a scalar,

$$(\alpha \mathbf{a}) \times \mathbf{b} = \alpha (\mathbf{a} \times \mathbf{b}) = \mathbf{a} \times (\alpha \mathbf{b}), \quad (14.16)$$

For \mathbf{a}, \mathbf{b} , and \mathbf{c} vectors, one obtains the distributive laws,

$$\mathbf{a} \times (\mathbf{b} + \mathbf{c}) = \mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c}, \quad (14.17)$$

$$(\mathbf{b} + \mathbf{c}) \times \mathbf{a} = \mathbf{b} \times \mathbf{a} + \mathbf{c} \times \mathbf{a}. \quad (14.18)$$

Formula 14.15 follows immediately from the definition. The vectors $\mathbf{a} \times \mathbf{b}$ and $\mathbf{b} \times \mathbf{a}$ have the same magnitude, $|\mathbf{a}||\mathbf{b}|\sin\theta$, and an application of the right hand rule shows they have opposite direction. Formula 14.16 is also fairly clear. If α is a nonnegative scalar, the direction of $(\alpha \mathbf{a}) \times \mathbf{b}$ is the same as the direction of $\mathbf{a} \times \mathbf{b}$, $\alpha (\mathbf{a} \times \mathbf{b})$ and $\mathbf{a} \times (\alpha \mathbf{b})$ while the magnitude is just α times the magnitude of $\mathbf{a} \times \mathbf{b}$ which is the same as the magnitude of $\alpha (\mathbf{a} \times \mathbf{b})$ and $\mathbf{a} \times (\alpha \mathbf{b})$. Using this yields equality in 14.16. In the case where $\alpha < 0$, everything works the same way except the vectors are all pointing in the opposite direction and you must multiply by $|\alpha|$ when comparing their magnitudes. The distributive laws are much harder to establish but the second follows from the first quite easily. Thus, assuming the first, and using 14.15,

$$(\mathbf{b} + \mathbf{c}) \times \mathbf{a} = -\mathbf{a} \times (\mathbf{b} + \mathbf{c}) = -(\mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c}) = \mathbf{b} \times \mathbf{a} + \mathbf{c} \times \mathbf{a}.$$

A proof of the distributive law is given later.

Now from the definition of the cross product,

$$\begin{aligned} \mathbf{i} \times \mathbf{j} &= \mathbf{k}, & \mathbf{j} \times \mathbf{i} &= -\mathbf{k} \\ \mathbf{k} \times \mathbf{i} &= \mathbf{j}, & \mathbf{i} \times \mathbf{k} &= -\mathbf{j} \\ \mathbf{j} \times \mathbf{k} &= \mathbf{i}, & \mathbf{k} \times \mathbf{j} &= -\mathbf{i} \end{aligned}$$

With this information, the following gives the coordinate description of the cross product.

Proposition 14.4.3 Let $\mathbf{a} = a_1\mathbf{i} + a_2\mathbf{j} + a_3\mathbf{k}$ and $\mathbf{b} = b_1\mathbf{i} + b_2\mathbf{j} + b_3\mathbf{k}$ be two vectors. Then

$$\mathbf{a} \times \mathbf{b} = (a_2b_3 - a_3b_2)\mathbf{i} + (a_3b_1 - a_1b_3)\mathbf{j} + (a_1b_2 - a_2b_1)\mathbf{k}. \quad (14.19)$$

Proof: From the above table and the properties of the cross product listed,

$$\begin{aligned}
 (a_1\mathbf{i} + a_2\mathbf{j} + a_3\mathbf{k}) \times (b_1\mathbf{i} + b_2\mathbf{j} + b_3\mathbf{k}) &= \\
 a_1b_2\mathbf{i} \times \mathbf{j} + a_1b_3\mathbf{i} \times \mathbf{k} + a_2b_1\mathbf{j} \times \mathbf{i} + a_2b_3\mathbf{j} \times \mathbf{k} + a_3b_1\mathbf{k} \times \mathbf{i} + a_3b_2\mathbf{k} \times \mathbf{j} \\
 &= a_1b_2\mathbf{k} - a_1b_3\mathbf{j} - a_2b_1\mathbf{k} + a_2b_3\mathbf{i} + a_3b_1\mathbf{j} - a_3b_2\mathbf{i} \\
 &= (a_2b_3 - a_3b_2)\mathbf{i} + (a_3b_1 - a_1b_3)\mathbf{j} + (a_1b_2 - a_2b_1)\mathbf{k} \quad (14.20)
 \end{aligned}$$

■

It is probably impossible for most people to remember 14.19. Fortunately, there is a somewhat easier way to remember it.

$$\mathbf{a} \times \mathbf{b} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix} \quad (14.21)$$

where you formally expand the determinant along the top row. For those who have not seen determinants, here is a short description. All you need here is how to evaluate 2×2 and 3×3 determinants.

$$\begin{vmatrix} x & y \\ z & w \end{vmatrix} = xw - yz$$

and

$$\begin{vmatrix} a & b & c \\ x & y & z \\ u & v & w \end{vmatrix} = a \begin{vmatrix} y & z \\ v & w \end{vmatrix} - b \begin{vmatrix} x & z \\ u & w \end{vmatrix} + c \begin{vmatrix} x & y \\ u & v \end{vmatrix}.$$

Here is the rule: You look at an entry in the top row and cross out the row and column which contain that entry. If the entry is in the i^{th} column, you multiply $(-1)^{1+i}$ times the determinant of the 2×2 which remains. This is the cofactor. You take the element in the top row times this cofactor and add all such terms. The rectangular array enclosed by the vertical lines is called a **matrix** and a lot more can be said about these, but this is enough for our purposes here.

Example 14.4.4 Find $(\mathbf{i} - \mathbf{j} + 2\mathbf{k}) \times (3\mathbf{i} - 2\mathbf{j} + \mathbf{k})$.

Use 14.21 to compute this.

$$\begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 1 & -1 & 2 \\ 3 & -2 & 1 \end{vmatrix} = \begin{vmatrix} -1 & 2 \\ -2 & 1 \end{vmatrix} \mathbf{i} - \begin{vmatrix} 1 & 2 \\ 3 & 1 \end{vmatrix} \mathbf{j} + \begin{vmatrix} 1 & -1 \\ 3 & -2 \end{vmatrix} \mathbf{k} = 3\mathbf{i} + 5\mathbf{j} + \mathbf{k}.$$

Example 14.4.5 Find the area of the parallelogram determined by the vectors

$$(\mathbf{i} - \mathbf{j} + 2\mathbf{k}), (3\mathbf{i} - 2\mathbf{j} + \mathbf{k}).$$

These are the same two vectors in Example 14.4.4.

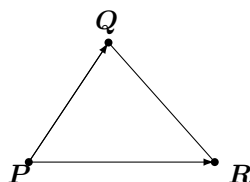
From Example 14.4.4 and the geometric description of the cross product, the area is just the norm of the vector obtained in Example 14.4.4. Thus the area is $\sqrt{9 + 25 + 1} = \sqrt{35}$.

Example 14.4.6 Find the area of the triangle determined by $(1, 2, 3)$, $(0, 2, 5)$, and $(5, 1, 2)$.

This triangle is obtained by connecting the three points with lines. Picking $(1, 2, 3)$ as a starting point, there are two displacement vectors $(-1, 0, 2)$ and $(4, -1, -1)$ such that the given vector added to these displacement vectors gives the other two vectors. The area of the triangle is half the area of the parallelogram determined by $(-1, 0, 2)$ and $(4, -1, -1)$. Thus $(-1, 0, 2) \times (4, -1, -1) = (2, 7, 1)$ and so the area of the triangle is $\frac{1}{2}\sqrt{4+49+1} = \frac{3}{2}\sqrt{6}$.

Observation 14.4.7 In general, if you have three points in \mathbb{R}^3 , P, Q, R the area of the triangle is given by

$$\frac{1}{2} |(Q - P) \times (R - P)|.$$



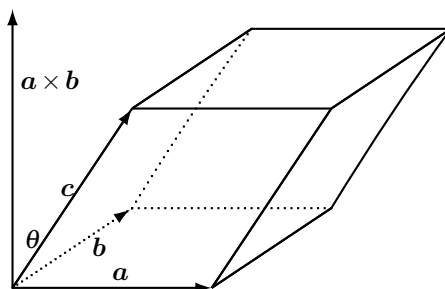
14.4.1 The Box Product

Definition 14.4.8 A parallelepiped determined by the three vectors \mathbf{a} , \mathbf{b} , and \mathbf{c} consists of

$$\{r\mathbf{a} + s\mathbf{b} + t\mathbf{c} : r, s, t \in [0, 1]\}.$$

That is, if you pick three numbers, r, s , and t each in $[0, 1]$ and form $r\mathbf{a} + s\mathbf{b} + t\mathbf{c}$, then the collection of all such points is what is meant by the parallelepiped determined by these three vectors.

The following is a picture of such a thing.



You notice the area of the base of the parallelepiped, the parallelogram determined by the vectors \mathbf{a} and \mathbf{b} has area equal to $|\mathbf{a} \times \mathbf{b}|$ while the altitude of the parallelepiped is $|\mathbf{c}| \cos \theta$ where θ is the angle shown in the picture between \mathbf{c} and $\mathbf{a} \times \mathbf{b}$. Therefore, the volume of this parallelepiped is the area of the base times the altitude which is just

$$|\mathbf{a} \times \mathbf{b}| |\mathbf{c}| \cos \theta = \mathbf{a} \times \mathbf{b} \cdot \mathbf{c}.$$

This expression is known as the box product and is sometimes written as $[\mathbf{a}, \mathbf{b}, \mathbf{c}]$. You should consider what happens if you interchange the \mathbf{b} with the \mathbf{c} or the \mathbf{a} with the \mathbf{c} . You

can see geometrically from drawing pictures that this merely introduces a minus sign. In any case the box product of three vectors always equals either the volume of the parallelepiped determined by the three vectors or else minus this volume.

Example 14.4.9 Find the volume of the parallelepiped determined by the vectors $\mathbf{i} + 2\mathbf{j} - 5\mathbf{k}$, $\mathbf{i} + 3\mathbf{j} - 6\mathbf{k}$, $3\mathbf{i} + 2\mathbf{j} + 3\mathbf{k}$.

According to the above discussion, pick any two of these, take the cross product and then take the dot product of this with the third of these vectors. The result will be either the desired volume or minus the desired volume.

$$(\mathbf{i} + 2\mathbf{j} - 5\mathbf{k}) \times (\mathbf{i} + 3\mathbf{j} - 6\mathbf{k}) = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 1 & 2 & -5 \\ 1 & 3 & -6 \end{vmatrix} = 3\mathbf{i} + \mathbf{j} + \mathbf{k}$$

Now take the dot product of this vector with the third which yields

$$(3\mathbf{i} + \mathbf{j} + \mathbf{k}) \cdot (3\mathbf{i} + 2\mathbf{j} + 3\mathbf{k}) = 9 + 2 + 3 = 14.$$

This shows the volume of this parallelepiped is 14 cubic units.

There is a fundamental observation which comes directly from the geometric definitions of the cross product and the dot product.

Lemma 14.4.10 Let \mathbf{a}, \mathbf{b} , and \mathbf{c} be vectors. Then $(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c} = \mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})$.

Proof: This follows from observing that either $(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c}$ and $\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})$ both give the volume of the parallelepiped or they both give -1 times the volume. ■

14.5 Proof of the Distributive Law

Let \mathbf{x} be a vector. From the above observation,

$$\begin{aligned} \mathbf{x} \cdot \mathbf{a} \times (\mathbf{b} + \mathbf{c}) &= (\mathbf{x} \times \mathbf{a}) \cdot (\mathbf{b} + \mathbf{c}) = (\mathbf{x} \times \mathbf{a}) \cdot \mathbf{b} + (\mathbf{x} \times \mathbf{a}) \cdot \mathbf{c} \\ &= \mathbf{x} \cdot \mathbf{a} \times \mathbf{b} + \mathbf{x} \cdot \mathbf{a} \times \mathbf{c} = \mathbf{x} \cdot (\mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c}). \end{aligned}$$

Therefore,

$$\mathbf{x} \cdot [\mathbf{a} \times (\mathbf{b} + \mathbf{c}) - (\mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c})] = 0$$

for all \mathbf{x} . In particular, this holds for $\mathbf{x} = \mathbf{a} \times (\mathbf{b} + \mathbf{c}) - (\mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c})$ showing that

$$\mathbf{a} \times (\mathbf{b} + \mathbf{c}) = \mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c}$$

and this proves the distributive law for the cross product.

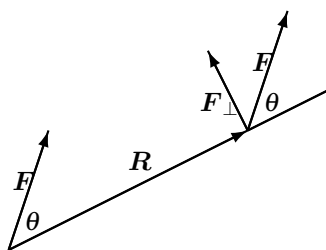
Observation 14.5.1 Suppose you have three vectors, $\mathbf{u} = (a, b, c)$, $\mathbf{v} = (d, e, f)$, and $\mathbf{w} = (g, h, i)$. Then $\mathbf{u} \cdot \mathbf{v} \times \mathbf{w}$ is given by the following.

$$\begin{aligned} \mathbf{u} \cdot \mathbf{v} \times \mathbf{w} &= (a, b, c) \cdot \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ d & e & f \\ g & h & i \end{vmatrix} = a \begin{vmatrix} e & f \\ h & i \end{vmatrix} - b \begin{vmatrix} d & f \\ g & i \end{vmatrix} + c \begin{vmatrix} d & e \\ g & h \end{vmatrix} \\ &= \det \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}. \end{aligned}$$

The message is that to take the box product, you can simply take the determinant of the matrix which results by letting the rows be the rectangular components of the given vectors in the order in which they occur in the box product.

14.5.1 Torque

Imagine you are using a wrench to loosen a nut. The idea is to turn the nut by applying a force to the end of the wrench. If you push or pull the wrench directly toward or away from the nut, it should be obvious from experience that no progress will be made in turning the nut. The important thing is the component of force perpendicular to the wrench. It is this component of force which will cause the nut to turn. For example see the following picture.



In the picture a force, F is applied at the end of a wrench represented by the position vector R and the angle between these two is θ . Then the tendency to turn will be $|R||F_{\perp}| = |R||F|\sin\theta$, which you recognize as the magnitude of the cross product of R and F . If there were just one force acting at one point whose position vector is R , perhaps this would be sufficient, but what if there are numerous forces acting at many different points with neither

the position vectors nor the force vectors in the same plane; what then? To keep track of this sort of thing, define for each R and F , the torque vector

$$\tau \equiv R \times F.$$

This is also called the moment of the force, F . That way, if there are several forces acting at several points the total torque can be obtained by simply adding up the torques associated with the different forces and positions.

Example 14.5.2 Suppose $R_1 = 2i - j + 3k$, $R_2 = i + 2j - 6k$ meters and at the points determined by these vectors there are forces, $F_1 = i - j + 2k$ and $F_2 = i - 5j + k$ Newtons respectively. Find the total torque about the origin produced by these forces acting at the given points.

It is necessary to take $R_1 \times F_1 + R_2 \times F_2$. Thus the total torque equals

$$\begin{vmatrix} i & j & k \\ 2 & -1 & 3 \\ 1 & -1 & 2 \end{vmatrix} + \begin{vmatrix} i & j & k \\ 1 & 2 & -6 \\ 1 & -5 & 1 \end{vmatrix} = -27i - 8j - 8k \text{ Newton meters}$$

Example 14.5.3 Find if possible a single force vector F which if applied at the point $i + j + k$ will produce the same torque as the above two forces acting at the given points.

This is fairly routine. The problem is to find $F = F_1i + F_2j + F_3k$ which produces the above torque vector. Therefore,

$$\begin{vmatrix} i & j & k \\ 1 & 1 & 1 \\ F_1 & F_2 & F_3 \end{vmatrix} = -27i - 8j - 8k$$

which reduces to $(F_3 - F_2)\mathbf{i} + (F_1 - F_3)\mathbf{j} + (F_2 - F_1)\mathbf{k} = -27\mathbf{i} - 8\mathbf{j} - 8\mathbf{k}$. This requirement amounts to solving the system of three equations in three unknowns, F_1, F_2 , and F_3 ,

$$F_3 - F_2 = -27, F_1 - F_3 = -8, F_2 - F_1 = -8$$

However, there is no solution to these three equations. (Why?) Therefore no single force acting at the point $\mathbf{i} + \mathbf{j} + \mathbf{k}$ will produce the given torque.

14.5.2 Center of Mass

The mass of an object is a measure of how much stuff there is in the object. An object has mass equal to one kilogram, a unit of mass in the metric system, if it would exactly balance a known one kilogram object when placed on a balance. The known object is one kilogram by definition. The mass of an object does not depend on where the balance is used. It would be one kilogram on the moon as well as on the earth. The weight of an object is something else. It is the force exerted on the object by gravity and has magnitude gm where g is a constant called the acceleration of gravity. Thus the weight of a one kilogram object would be different on the moon which has much less gravity, smaller g , than on the earth. An important idea is that of the center of mass. This is the point at which an object will balance no matter how it is turned.

Definition 14.5.4 *Let an object consist of p point masses m_1, \dots, m_p with the position of the k^{th} of these at \mathbf{R}_k . The center of mass of this object \mathbf{R}_0 is the point satisfying*

$$\sum_{k=1}^p (\mathbf{R}_k - \mathbf{R}_0) \times g m_k \mathbf{u} = \mathbf{0}$$

for all unit vectors \mathbf{u} .

The above definition indicates that no matter how the object is suspended, the total torque on it due to gravity is such that no rotation occurs. Using the properties of the cross product

$$\left(\sum_{k=1}^p \mathbf{R}_k g m_k - \mathbf{R}_0 \sum_{k=1}^p g m_k \right) \times \mathbf{u} = \mathbf{0} \quad (14.22)$$

for any choice of unit vector \mathbf{u} . You should verify that if $\mathbf{a} \times \mathbf{u} = \mathbf{0}$ for all \mathbf{u} , then it must be the case that $\mathbf{a} = \mathbf{0}$. Then the above formula requires that

$$\sum_{k=1}^p \mathbf{R}_k g m_k - \mathbf{R}_0 \sum_{k=1}^p g m_k = \mathbf{0}.$$

dividing by g , and then by $\sum_{k=1}^p m_k$,

$$\mathbf{R}_0 = \frac{\sum_{k=1}^p \mathbf{R}_k m_k}{\sum_{k=1}^p m_k}. \quad (14.23)$$

This is the formula for the center of mass of a collection of point masses. To consider the center of mass of a solid consisting of continuously distributed masses, you need the methods of multi-variable calculus.

Example 14.5.5 Let $m_1 = 5$, $m_2 = 6$, and $m_3 = 3$ where the masses are in kilograms. Suppose m_1 is located at $2\mathbf{i} + 3\mathbf{j} + \mathbf{k}$, m_2 is located at $\mathbf{i} - 3\mathbf{j} + 2\mathbf{k}$ and m_3 is located at $2\mathbf{i} - \mathbf{j} + 3\mathbf{k}$. Find the center of mass of these three masses.

Using 14.23

$$\mathbf{R}_0 = \frac{5(2\mathbf{i} + 3\mathbf{j} + \mathbf{k}) + 6(\mathbf{i} - 3\mathbf{j} + 2\mathbf{k}) + 3(2\mathbf{i} - \mathbf{j} + 3\mathbf{k})}{5 + 6 + 3} = \frac{11}{7}\mathbf{i} - \frac{3}{7}\mathbf{j} + \frac{13}{7}\mathbf{k}$$

14.5.3 Angular Velocity

Definition 14.5.6 In a rotating body, a vector $\boldsymbol{\Omega}$ is called an **angular velocity vector** if the velocity of a point having position vector \mathbf{u} relative to the body is given by $\boldsymbol{\Omega} \times \mathbf{u}$.

The existence of an angular velocity vector is the key to understanding motion in a moving system of coordinates. It is used to explain the motion on the surface of the rotating earth. For example, have you ever wondered why low pressure areas rotate counter clockwise in the Northern hemisphere but clockwise in the Southern hemisphere? To quantify these things, you will need the concept of an angular velocity vector. Here is a simple example. Think of a coordinate system fixed in the rotating body. Thus if you were riding on the rotating body, you would observe this coordinate system as fixed.

Example 14.5.7 A wheel rotates counter clockwise about the vector $\mathbf{i} + \mathbf{j} + \mathbf{k}$ at 60 revolutions per minute. This means that if the thumb of your right hand were to point in the direction of $\mathbf{i} + \mathbf{j} + \mathbf{k}$ your fingers of this hand would wrap in the direction of rotation. Find the angular velocity vector for this wheel. Assume the unit of distance is meters and the unit of time is minutes.

Let $\omega = 60 \times 2\pi = 120\pi$. This is the number of radians per minute corresponding to 60 revolutions per minute. Then the angular velocity vector is $\frac{120\pi}{\sqrt{3}}(\mathbf{i} + \mathbf{j} + \mathbf{k})$. Note this gives what you would expect in the case the position vector to the point is perpendicular to $\mathbf{i} + \mathbf{j} + \mathbf{k}$ and at a distance of r . This is because of the geometric description of the cross product. The magnitude of the vector is $r120\pi$ meters per minute and corresponds to the speed and an exercise with the right hand shows the direction is correct also. However, if this body is rigid, this will work for every other point in it, even those for which the position vector is not perpendicular to the given vector.

Example 14.5.8 A wheel rotates counter clockwise about the vector $\mathbf{i} + \mathbf{j} + \mathbf{k}$ at 60 revolutions per minute exactly as in Example 14.5.7. Let $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ denote an orthogonal right handed system attached to the rotating wheel in which $\mathbf{u}_3 = \frac{1}{\sqrt{3}}(\mathbf{i} + \mathbf{j} + \mathbf{k})$. Thus \mathbf{u}_1 and \mathbf{u}_2 depend on time but, $\mathbf{u}_1 \times \mathbf{u}_2 = \mathbf{u}_3$. Find the velocity of the point of the wheel located at the point $2\mathbf{u}_1 + 3\mathbf{u}_2 - \mathbf{u}_3$. Note this point is not fixed in space. It is moving.

Since $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ is a right handed system like $\mathbf{i}, \mathbf{j}, \mathbf{k}$, everything applies to this system in the same way as with $\mathbf{i}, \mathbf{j}, \mathbf{k}$. Thus the cross product is given by

$$(a\mathbf{u}_1 + b\mathbf{u}_2 + c\mathbf{u}_3) \times (d\mathbf{u}_1 + e\mathbf{u}_2 + f\mathbf{u}_3) = \begin{vmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 \\ a & b & c \\ d & e & f \end{vmatrix}$$

Therefore, in terms of the given vectors \mathbf{u}_i , the angular velocity vector is $120\pi\mathbf{u}_3$. The velocity of the given point is

$$\begin{vmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 \\ 0 & 0 & 120\pi \\ 2 & 3 & -1 \end{vmatrix} = -360\pi\mathbf{u}_1 + 240\pi\mathbf{u}_2$$

in meters per minute. Note how this gives the answer in terms of these vectors which are fixed in the body, not in space. Since \mathbf{u}_i depends on t , this shows the answer in this case does also. Of course this is right. Just think of what is going on with the wheel rotating. Those vectors which are fixed in the wheel are moving in space relative to a stationary observer. The velocity of a point in the wheel should be constantly changing. However, its speed will not change. The speed will be the magnitude of the velocity and this is

$$\sqrt{(-360\pi\mathbf{u}_1 + 240\pi\mathbf{u}_2) \cdot (-360\pi\mathbf{u}_1 + 240\pi\mathbf{u}_2)}$$

which from the properties of the dot product equals

$$\sqrt{(-360\pi)^2 + (240\pi)^2} = 120\sqrt{13}\pi$$

because the \mathbf{u}_i are given to be orthogonal.

14.6 Vector Identities and Notation

To begin with consider $\mathbf{u} \times (\mathbf{v} \times \mathbf{w})$ and it is desired to simplify this expression. It turns out this expression comes up in many different contexts. Let $\mathbf{u} = (u_1, u_2, u_3)$ and let \mathbf{v} and \mathbf{w} be defined similarly.

$$\mathbf{v} \times \mathbf{w} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix} = (v_2w_3 - v_3w_2)\mathbf{i} + (w_1v_3 - v_1w_3)\mathbf{j} + (v_1w_2 - v_2w_1)\mathbf{k}$$

Next consider $\mathbf{u} \times (\mathbf{v} \times \mathbf{w})$ which is given by

$$\mathbf{u} \times (\mathbf{v} \times \mathbf{w}) = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ u_1 & u_2 & u_3 \\ (v_2w_3 - v_3w_2) & (w_1v_3 - v_1w_3) & (v_1w_2 - v_2w_1) \end{vmatrix}.$$

When you multiply this out, you get

$$\begin{aligned} & \mathbf{i}(v_1u_2w_2 + u_3v_1w_3 - w_1u_2v_2 - u_3w_1v_3) + \mathbf{j}(v_2u_1w_1 + v_2w_3u_3 - w_2u_1v_1 - u_3w_2v_3) \\ & + \mathbf{k}(u_1w_1v_3 + v_3w_2u_2 - u_1v_1w_3 - v_2w_3u_2) \end{aligned}$$

and if you are clever, you see right away that

$$(\mathbf{i}v_1 + \mathbf{j}v_2 + \mathbf{k}v_3)(u_1w_1 + u_2w_2 + u_3w_3) - (\mathbf{i}w_1 + \mathbf{j}w_2 + \mathbf{k}w_3)(u_1v_1 + u_2v_2 + u_3v_3).$$

Thus

$$\mathbf{u} \times (\mathbf{v} \times \mathbf{w}) = \mathbf{v}(\mathbf{u} \cdot \mathbf{w}) - \mathbf{w}(\mathbf{u} \cdot \mathbf{v}). \quad (14.24)$$

A related formula is

$$\begin{aligned} (\mathbf{u} \times \mathbf{v}) \times \mathbf{w} &= -[\mathbf{w} \times (\mathbf{u} \times \mathbf{v})] = -[\mathbf{u}(\mathbf{w} \cdot \mathbf{v}) - \mathbf{v}(\mathbf{w} \cdot \mathbf{u})] \\ &= \mathbf{v}(\mathbf{w} \cdot \mathbf{u}) - \mathbf{u}(\mathbf{w} \cdot \mathbf{v}). \end{aligned} \quad (14.25)$$

This derivation is simply wretched and it does nothing for other identities which may arise in applications. Actually, the above two formulas, 14.24 and 14.25 are sufficient for most applications if you are creative in using them, but there is another way. This other way allows you to discover such vector identities as the above without any creativity or any cleverness. Therefore, it is far superior to the above nasty and tedious computation. It is a vector identity discovering machine and it is this which is the main topic in what follows. I cannot understand why it is not routinely presented in calculus texts. The engineers I have known seem to know all about it.

There are two special symbols, δ_{ij} and ϵ_{ijk} which are very useful in dealing with vector identities. To begin with, here is the definition of these symbols.

Definition 14.6.1 The symbol δ_{ij} , called the Kronecker delta symbol is defined as follows.

$$\delta_{ij} \equiv \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}.$$

With the Kronecker symbol i and j can equal any integer in $\{1, 2, \dots, n\}$ for any $n \in \mathbb{N}$.

Definition 14.6.2 For i, j , and k integers in the set, $\{1, 2, 3\}$, ϵ_{ijk} is defined as follows.

$$\epsilon_{ijk} \equiv \begin{cases} 1 & \text{if } (i, j, k) = (1, 2, 3), (2, 3, 1), \text{ or } (3, 1, 2) \\ -1 & \text{if } (i, j, k) = (2, 1, 3), (1, 3, 2), \text{ or } (3, 2, 1) \\ 0 & \text{if there are any repeated integers} \end{cases}.$$

The subscripts ijk and ij in the above are called indices. A single one is called an index. This symbol ϵ_{ijk} is also called the permutation symbol.

The way to think of ϵ_{ijk} is that $\epsilon_{123} = 1$ and if you switch any two of the numbers in the list i, j, k , it changes the sign. Thus $\epsilon_{ijk} = -\epsilon_{jik}$ and $\epsilon_{ijk} = -\epsilon_{kji}$ etc. You should check that this rule reduces to the above definition. For example, it immediately implies that if there is a repeated index, the answer is zero. This follows because $\epsilon_{iij} = -\epsilon_{iij}$ and so $\epsilon_{iij} = 0$.

It is useful to use the Einstein summation convention when dealing with these symbols. Simply stated, the convention is that you sum over the repeated index. Thus $a_i b_i$ means $\sum_i a_i b_i$. Also, $\delta_{ij} x_j$ means $\sum_j \delta_{ij} x_j = x_i$. When you use this convention, there is one very important thing to never forget. It is this: Never have an index be repeated more than once. Thus $a_i b_i$ is all right but $a_{ii} b_i$ is not. The reason for this is that you end up getting confused about what is meant. If you want to write $\sum_i a_i b_i c_i$ it is best to simply use the summation notation. There is a very important reduction identity connecting these two symbols.

Lemma 14.6.3 The following holds.

$$\epsilon_{ijk} \epsilon_{irs} = (\delta_{jr} \delta_{ks} - \delta_{kr} \delta_{js}).$$

Proof: If $\{j, k\} \neq \{r, s\}$ then every term in the sum on the left must have either ϵ_{ijk} or ϵ_{irs} contains a repeated index. Therefore, the left side equals zero. The right side also equals zero in this case. To see this, note that if the two sets of indices are not equal, then

there is one of the indices in one of the sets which is not in the other set. For example, it could be that j is not equal to either r or s . Then the right side equals zero.

Therefore, it can be assumed $\{j, k\} = \{r, s\}$. If $i = r$ and $j = s$ for $s \neq r$, then there is exactly one term in the sum on the left and it equals 1. The right also reduces to 1 in this case. If $i = s$ and $j = r$, there is exactly one term in the sum on the left which is nonzero and it must equal -1 . The right side also reduces to -1 in this case. If there is a repeated index in $\{j, k\}$, then every term in the sum on the left equals zero. The right also reduces to zero in this case because then $j = k = r = s$ and so the right side becomes $(1)(1) - (-1)(-1) = 0$. ■

Proposition 14.6.4 *Let \mathbf{u}, \mathbf{v} be vectors in \mathbb{R}^p where the Cartesian coordinates of \mathbf{u} are (u_1, \dots, u_p) and the Cartesian coordinates of \mathbf{v} are (v_1, \dots, v_p) . Then $\mathbf{u} \cdot \mathbf{v} = u_i v_i$. If \mathbf{u}, \mathbf{v} are vectors in \mathbb{R}^3 , then*

$$(\mathbf{u} \times \mathbf{v})_i = \varepsilon_{ijk} u_j v_k.$$

Also, $\delta_{ik} a_k = a_i$.

Proof: The first claim is obvious from the definition of the dot product. The second is verified by simply checking it works. For example,

$$\mathbf{u} \times \mathbf{v} \equiv \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix}$$

and so

$$(\mathbf{u} \times \mathbf{v})_1 = (u_2 v_3 - u_3 v_2).$$

From the above formula in the proposition,

$$\varepsilon_{1jk} u_j v_k \equiv u_2 v_3 - u_3 v_2,$$

the same thing. The cases for $(\mathbf{u} \times \mathbf{v})_2$ and $(\mathbf{u} \times \mathbf{v})_3$ are verified similarly. The last claim follows directly from the definition. ■

With this notation, you can easily discover vector identities and simplify expressions which involve the cross product.

Example 14.6.5 *Discover a formula which simplifies $(\mathbf{u} \times \mathbf{v}) \times \mathbf{w}$.*

From the above reduction formula,

$$\begin{aligned} ((\mathbf{u} \times \mathbf{v}) \times \mathbf{w})_i &= \varepsilon_{ijk} (\mathbf{u} \times \mathbf{v})_j w_k = \varepsilon_{ijk} \varepsilon_{jrs} u_r v_s w_k \\ &= -\varepsilon_{jik} \varepsilon_{jrs} u_r v_s w_k = -(\delta_{ir} \delta_{ks} - \delta_{is} \delta_{kr}) u_r v_s w_k \\ &= -(u_i v_k w_k - u_k v_i w_k) = \mathbf{u} \cdot \mathbf{w} v_i - \mathbf{v} \cdot \mathbf{w} u_i \\ &= ((\mathbf{u} \cdot \mathbf{w}) \mathbf{v} - (\mathbf{v} \cdot \mathbf{w}) \mathbf{u})_i. \end{aligned}$$

Since this holds for all i , it follows that

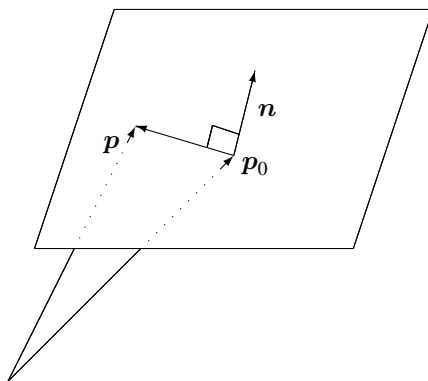
$$(\mathbf{u} \times \mathbf{v}) \times \mathbf{w} = (\mathbf{u} \cdot \mathbf{w}) \mathbf{v} - (\mathbf{v} \cdot \mathbf{w}) \mathbf{u}.$$

14.7 Planes

This section concerns something called a level surface of a function of many variables. It is a little outside the goals of this book but it seems a shame not to consider it because it is a nice illustration of the geometric significance of the dot product. To find the equation of a plane, you need two things, a point contained in the plane and a vector normal to the plane. Let $\mathbf{p}_0 = (x_0, y_0, z_0)$ denote the position vector of a point in the plane, let $\mathbf{p} = (x, y, z)$ be the position vector of an arbitrary point in the plane, and let \mathbf{n} denote a vector normal to the plane. This means that

$$\mathbf{n} \cdot (\mathbf{p} - \mathbf{p}_0) = 0$$

whenever \mathbf{p} is the position vector of a point in the plane. The following picture illustrates the geometry of this idea.



Expressed equivalently, the plane is just the set of all points \mathbf{p} such that the vector $\mathbf{p} - \mathbf{p}_0$ is perpendicular to the given normal vector \mathbf{n} .

Example 14.7.1 Find the equation of the plane with normal vector $\mathbf{n} = (1, 2, 3)$ containing the point $(2, -1, 5)$.

From the above, the equation of this plane is just

$$(1, 2, 3) \cdot (x - 2, y + 1, z - 5) = 0 \quad \text{or} \quad x - 9 + 2y + 3z = 0$$

Example 14.7.2 $2x + 4y - 5z = 11$ is the equation of a plane. Find the normal vector and a point on this plane.

You can write this in the form $2(x - \frac{11}{2}) + 4(y - 0) + (-5)(z - 0) = 0$. Therefore, a normal vector to the plane is $2\mathbf{i} + 4\mathbf{j} - 5\mathbf{k}$ and a point in this plane is $(\frac{11}{2}, 0, 0)$. Of course there are many other points in the plane.

Definition 14.7.3 Suppose two planes intersect. The angle between the planes is defined to be the angle which is less than $\pi/2$ between normal vectors to the respective planes.

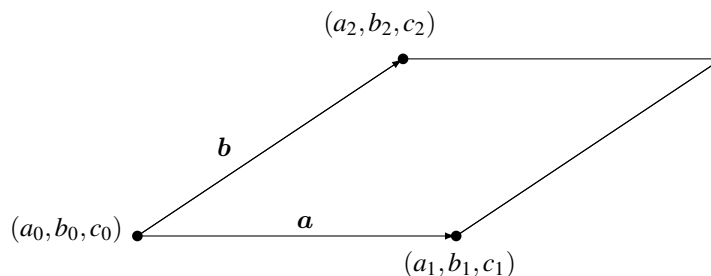
Example 14.7.4 Find the angle between the two planes $x + 2y - z = 6$ and $3x + 2y - z = 7$.

The two normal vectors are $(1, 2, -1)$ and $(3, 2, -1)$. Therefore, the cosine of the angle desired is

$$\cos \theta = \frac{(1, 2, -1) \cdot (3, 2, -1)}{\sqrt{1^2 + 2^2 + (-1)^2} \sqrt{3^2 + 2^2 + (-1)^2}} = .87287$$

Now use a calculator or table to find what the angle is. $\cos \theta = .87287$, Solution is : $\{\theta = .50974\}$. This value is in radians.

Sometimes you need to find the equation of a plane which contains three points. Consider the following picture.



You have plenty of points but you need a normal. This can be obtained by taking $\mathbf{a} \times \mathbf{b}$ where $\mathbf{a} = (a_1 - a_0, b_1 - b_0, c_1 - c_0)$ and $\mathbf{b} = (a_2 - a_0, b_2 - b_0, c_2 - c_0)$.

Example 14.7.5 Find the equation of the plane which contains the three points

$$(1, 2, 1), (3, -1, 2), \text{ and } (4, 2, 1).$$

You just need to get a normal vector to this plane. This can be done by taking the cross products of the two vectors

$$(3, -1, 2) - (1, 2, 1) \text{ and } (4, 2, 1) - (1, 2, 1)$$

Thus a normal vector is $(2, -3, 1) \times (3, 0, 0) = (0, 3, 9)$. Therefore, the equation of the plane is

$$0(x - 1) + 3(y - 2) + 9(z - 1) = 0$$

or $3y + 9z = 15$ which is the same as $y + 3z = 5$. When you have what you think is the plane containing the three points, you ought to check it by seeing if it really does contain the three points.

Proposition 14.7.6 If $(a, b, c) \neq (0, 0, 0)$, then $ax + by + cz = d$ is the equation of a plane with normal vector $a\mathbf{i} + b\mathbf{j} + c\mathbf{k}$. Conversely, any plane can be written in this form.

Proof: One of a, b, c is nonzero. Suppose for example that $c \neq 0$. Then the equation can be written as

$$a(x - 0) + b(y - 0) + c\left(z - \frac{d}{c}\right) = 0$$

Therefore, $(0, 0, \frac{d}{c})$ is a point on the plane and a normal vector is $a\mathbf{i} + b\mathbf{j} + c\mathbf{k}$. The converse follows from the above discussion involving the point and a normal vector. ■

Example 14.7.7 Find the equation of the plane containing the points $(1, 2, 3)$ and the line $(0, 1, 1) + t(2, 1, 2) = (x, y, z)$.

There are several ways to do this. One is to find three points and use the above procedures. Let $t = 0$ and then let $t = 1$ to get two points on the line. This yields the three points $(1, 2, 3)$, $(0, 1, 1)$, and $(2, 2, 3)$. Then a normal vector is obtained by fixing a point and taking the cross product of the differences of the other two points with that one. Thus in this case, fixing $(0, 1, 1)$, a normal vector is

$$(1, 1, 2) \times (2, 1, 2) = (0, 2, -1)$$

Therefore, an equation for the plane is

$$0(x - 0) + 2(y - 1) + (-1)(x - 3) = 0$$

Simplifying this yields

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

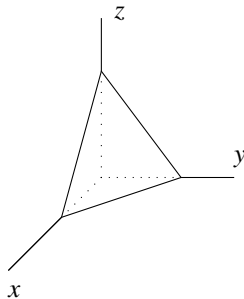
Example 14.7.8 Find the equation of the plane which contains the two lines, given by the following parametric expressions in which $t \in \mathbb{R}$.

$$(2t, 1 + t, 1 + 2t) = (x, y, z), \quad (2t + 2, 1, 3 + 2t) = (x, y, z)$$

Note first that you don't know there even is such a plane. However, if there is, you could find it by obtaining three points, two on one line and one on another and then using any of the above procedures for finding the plane. From the first line, two points are $(0, 1, 1)$ and $(2, 2, 3)$ while a third point can be obtained from second line, $(2, 1, 3)$. You need a normal vector and then use any of these points. To get a normal vector, form $(2, 0, 2) \times (2, 1, 2) = (-2, 0, 2)$. Therefore, the plane is $-2x + 0(y - 1) + 2(z - 1) = 0$. This reduces to $z - x = 1$. If there is a plane, this is it. Now you can simply verify that both of the lines are really in this plane. From the first, $(1 + 2t) - 2t = 1$ and the second, $(3 + 2t) - (2t + 2) = 1$ so both lines lie in the plane.

One way to understand how a plane looks is to connect the points where it intercepts the x , y , and z axes. This allows you to visualize the plane somewhat and is a good way to sketch the plane. Not surprisingly these points are called intercepts.

Example 14.7.9 Sketch the plane having intercepts $(2, 0, 0)$, $(0, 3, 0)$, and $(0, 0, 4)$.



You see how connecting the intercepts gives a fairly good geometric description of the plane. These lines which connect the intercepts are also called the traces of the plane. Thus the line which joins $(0, 3, 0)$ to $(0, 0, 4)$ is the intersection of the plane with the yz plane. It is the trace on the yz plane.

Example 14.7.10 Identify the intercepts of the plane $3x - 4y + 5z = 11$.

The easy way to do this is to divide both sides by 11. Thus $\frac{x}{(11/3)} + \frac{y}{(-11/4)} + \frac{z}{(11/5)} = 1$. The intercepts are $(11/3, 0, 0)$, $(0, -11/4, 0)$ and $(0, 0, 11/5)$. You can see this by letting both y and z equal to zero to find the point on the x axis which is intersected by the plane. The other axes are handled similarly.

14.8 Exercises

1. Show that if $\mathbf{a} \times \mathbf{u} = \mathbf{0}$ for all unit vectors \mathbf{u} , then $\mathbf{a} = \mathbf{0}$.
2. If you only assume 14.22 holds for $\mathbf{u} = \mathbf{i}, \mathbf{j}, \mathbf{k}$, show that this implies 14.22 holds for all unit vectors \mathbf{u} .
3. Let $m_1 = 5, m_2 = 1$, and $m_3 = 4$ where the masses are in kilograms and the distance is in meters. Suppose m_1 is located at $2\mathbf{i} - 3\mathbf{j} + \mathbf{k}$, m_2 is located at $\mathbf{i} - 3\mathbf{j} + 6\mathbf{k}$ and m_3 is located at $2\mathbf{i} + \mathbf{j} + 3\mathbf{k}$. Find the center of mass of these three masses.
4. Let $m_1 = 2, m_2 = 3$, and $m_3 = 1$ where the masses are in kilograms and the distance is in meters. Suppose m_1 is located at $2\mathbf{i} - \mathbf{j} + \mathbf{k}$, m_2 is located at $\mathbf{i} - 2\mathbf{j} + \mathbf{k}$ and m_3 is located at $4\mathbf{i} + \mathbf{j} + 3\mathbf{k}$. Find the center of mass of these three masses.
5. Find the angular velocity vector of a rigid body which rotates counter clockwise about the vector $\mathbf{i} - 2\mathbf{j} + \mathbf{k}$ at 40 revolutions per minute. Assume distance is measured in meters.
6. Let $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ be a right handed system with \mathbf{u}_3 pointing in the direction of $\mathbf{i} - 2\mathbf{j} + \mathbf{k}$ and \mathbf{u}_1 and \mathbf{u}_2 being fixed with the body which is rotating at 40 revolutions per minute. This is also an orthonormal system meaning $\mathbf{u}_i \cdot \mathbf{u}_j = \delta_{ij}$. Assuming all distances are in meters, find the constant speed of the point of the body located at $3\mathbf{u}_1 + \mathbf{u}_2 - \mathbf{u}_3$ in meters per minute.
7. Find the area of the triangle determined by the three points $(1, 2, 3)$, $(4, 2, 0)$ and $(-3, 2, 1)$.
8. Find the area of the triangle determined by the three points $(1, 2, 3)$, $(2, 3, 4)$ and $(0, 1, 2)$. Did something interesting happen here? What does it mean geometrically?
9. Find the area of the parallelogram determined by the vectors $(1, 2, 3)$ and $(3, -2, 1)$.
10. Find the area of the parallelogram determined by the vectors $(1, -2, 2)$ and $(3, 1, 1)$.
11. Find the volume of the parallelepiped determined by the vectors $\mathbf{i} - 7\mathbf{j} - 5\mathbf{k}$, $\mathbf{i} - 2\mathbf{j} - 6\mathbf{k}$, $3\mathbf{i} + 2\mathbf{j} + 3\mathbf{k}$.
12. Find the volume of the parallelepiped determined by the vectors $\mathbf{i} + \mathbf{j} - 5\mathbf{k}$, $\mathbf{i} + 5\mathbf{j} - 6\mathbf{k}$, $3\mathbf{i} + \mathbf{j} + 3\mathbf{k}$.
13. Find the volume of the parallelepiped determined by the vectors $\mathbf{i} + 6\mathbf{j} + 5\mathbf{k}$, $\mathbf{i} + 5\mathbf{j} - 6\mathbf{k}$, $3\mathbf{i} + \mathbf{j} + \mathbf{k}$.
14. Suppose \mathbf{a}, \mathbf{b} , and \mathbf{c} are three vectors whose components are all integers. Can you conclude the volume of the parallelepiped determined from these three vectors will always be an integer?

15. What does it mean geometrically if the box product of three vectors gives zero?
16. It is desired to find an equation of a plane parallel to the two vectors \mathbf{a} and \mathbf{b} containing the point $\mathbf{0}$. Using Problem 15, show an equation for this plane is

$$\begin{vmatrix} x & y & z \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix} = 0$$

That is, the set of all (x, y, z) such that

$$x \begin{vmatrix} a_2 & a_3 \\ b_2 & b_3 \end{vmatrix} - y \begin{vmatrix} a_1 & a_3 \\ b_1 & b_3 \end{vmatrix} + z \begin{vmatrix} a_1 & a_2 \\ b_1 & b_2 \end{vmatrix} = 0$$

17. Using the notion of the box product yielding either plus or minus the volume of the parallelepiped determined by the given three vectors, show that

$$(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c} = \mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})$$

In other words, the dot and the cross can be switched as long as the order of the vectors remains the same. **Hint:** There are two ways to do this, by the coordinate description of the dot and cross product and by geometric reasoning.

18. Is $\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = (\mathbf{a} \times \mathbf{b}) \times \mathbf{c}$? What is the meaning of $\mathbf{a} \times \mathbf{b} \times \mathbf{c}$? Explain. **Hint:** Try $(\mathbf{i} \times \mathbf{j}) \times \mathbf{j}$.
19. Verify directly that the coordinate description of the cross product $\mathbf{a} \times \mathbf{b}$ has the property that it is perpendicular to both \mathbf{a} and \mathbf{b} . Then show by direct computation that this coordinate description satisfies

$$|\mathbf{a} \times \mathbf{b}|^2 = |\mathbf{a}|^2 |\mathbf{b}|^2 - (\mathbf{a} \cdot \mathbf{b})^2 = |\mathbf{a}|^2 |\mathbf{b}|^2 (1 - \cos^2(\theta))$$

where θ is the angle included between the two vectors. Explain why $|\mathbf{a} \times \mathbf{b}|$ has the correct magnitude. All that is missing is the material about the right hand rule. Verify directly that the right thing happens with regards to the vectors $\mathbf{i}, \mathbf{j}, \mathbf{k}$. Next verify that the distributive law holds for the coordinate description of the cross product. This gives another way to approach the cross product. First define it in terms of coordinates and then get the geometric properties from this.

20. Discover a vector identity for $\mathbf{u} \times (\mathbf{v} \times \mathbf{w})$.
21. Discover a vector identity for $(\mathbf{u} \times \mathbf{v}) \cdot (\mathbf{z} \times \mathbf{w})$.
22. Discover a vector identity for $(\mathbf{u} \times \mathbf{v}) \times (\mathbf{z} \times \mathbf{w})$ in terms of box products.
23. Simplify $(\mathbf{u} \times \mathbf{v}) \cdot (\mathbf{v} \times \mathbf{w}) \times (\mathbf{w} \times \mathbf{z})$.
24. Simplify $|\mathbf{u} \times \mathbf{v}|^2 + (\mathbf{u} \cdot \mathbf{v})^2 - |\mathbf{u}|^2 |\mathbf{v}|^2$.
25. Prove that $\epsilon_{ijk} \epsilon_{ijr} = 2\delta_{kr}$.
26. If A is a 3×3 matrix such that $A = \begin{pmatrix} \mathbf{u} & \mathbf{v} & \mathbf{w} \end{pmatrix}$ where these are the columns of the matrix A . Show that $\det(A) = \epsilon_{ijk} u_i v_j w_k$.

27. If A is a 3×3 matrix, show $\epsilon_{rps} \det(A) = \epsilon_{ijk} A_{ri} A_{pj} A_{sk}$.
28. Suppose A is a 3×3 matrix and $\det(A) \neq 0$. Show using 27 and 25 that

$$(A^{-1})_{ks} = \frac{1}{\det(A)} \epsilon_{rps} \epsilon_{ijk} A_{pj} A_{ri}.$$

29. When you have a rotating rigid body with angular velocity vector Ω then the velocity, \mathbf{u}' is given by $\mathbf{u}' = \Omega \times \mathbf{u}$. It turns out that all the usual calculus rules such as the product rule hold. Also, \mathbf{u}'' is the acceleration. Show using the product rule that for Ω a constant vector

$$\mathbf{u}'' = \Omega \times (\Omega \times \mathbf{u}).$$

It turns out this is the centripetal acceleration. Note how it involves cross products.

30. Find the planes which go through the following collections of three points. In case the plane is not well defined, explain why.
- (a) $(1, 2, 0), (2, -1, 1), (3, 1, 1)$
 - (b) $(3, 1, 0), (2, 1, 1), (-3, 1, -1)$
 - (c) $(2, 1, 1), (-2, 3, 1), (0, 4, 2)$
 - (d) $(1, 0, 1), (2, 0, 1), (0, 1, 1)$
31. A point is given along with a line. Find the equation for the plane which contains the line as well as the point.
- (a) $(1, 2, 1), (1, -1, 1) + t(1, 0, 1)$
 - (b) $(2, 1, -1), (1, 1, 1) + t(2, -1, 1)$
 - (c) $(-1, 2, 3), (-1, 1, 1) + t(2, 1, 1)$
 - (d) $(2, 0, 1), (2, 1, 1) + t(-1, 1, 1)$

Chapter 15

Sequences, Compactness, and Continuity

This chapter is on open, closed, and compact sets in \mathbb{R}^p . The reason this is done is to show fundamental results about continuous functions which will also be defined a little later.

15.1 Sequences of Vectors

Recall how a sequence was a function from a set $\{m, m+1, \dots\}$ to \mathbb{R} . This was the case discussed earlier anyway. It is no different if the function has values in \mathbb{R}^p . A vector valued sequence is just a function from $\{m, m+1, \dots\}$ with values in \mathbb{R}^p . Convergence of sequences is defined exactly as before. Note that saying $|\mathbf{x} - \mathbf{y}|$ is small is exactly the same as saying that $|x_i - y_i|$ is small for each i . This is easily seen by observing that

$$\begin{aligned} \max \{|x_i - y_i| : i \leq p\} &\leq |\mathbf{x} - \mathbf{y}| = \sqrt{\sum_{i=1}^p |x_i - y_i|^2} \\ &\leq \sqrt{\sum_{i=1}^p \max \{|x_i - y_i|^2 : i \leq p\}} = \sqrt{p} \max \{|x_i - y_i| : i \leq p\} \end{aligned} \quad (15.1)$$

Definition 15.1.1 $\lim_{n \rightarrow \infty} \mathbf{x}^n = \mathbf{x}$ means: For every $\varepsilon > 0$ there is n_ε such that if $n \geq n_\varepsilon$, then $|\mathbf{x}^n - \mathbf{x}| < \varepsilon$. From the description of $|\cdot|$ given earlier, this says the same as $\lim_{n \rightarrow \infty} x_i^n = x_i$ for each $i = 1, 2, \dots, p$ where $\mathbf{x}^n \equiv (x_1^n, \dots, x_p^n)$ and $\mathbf{x} \equiv (x_1, \dots, x_p)$.

As just explained, there isn't a lot new here. Convergence of a sequence of vectors is equivalent to consideration of convergence of the components of the sequence.

Also similar is the concept of a Cauchy sequence.

Definition 15.1.2 $\{\mathbf{x}_k\}$ is a Cauchy sequence if and only if the following holds. For every $\varepsilon > 0$, there exists n_ε such that if $k, l \geq n_\varepsilon$, then $|\mathbf{x}_k - \mathbf{x}_l| < \varepsilon$.

As explained above, a sequence $\{\mathbf{x}_k\}$ is Cauchy if and only if the sequences of components of $\{\mathbf{x}_k\}$ are Cauchy sequences. The following theorem follows from this.

Theorem 15.1.3 A sequence $\{x^k\}$ converges if and only if it is a Cauchy sequence.

Proof: Let $x^k = (x_1^k, \dots, x_p^k)$. Then from 15.1, $\{x^k\}$ is Cauchy if and only if $\{x_i^k\}_{k=1}^\infty$ is Cauchy for each $i \leq p$ if and only if $\{x_i^k\}_{k=1}^\infty$ converges to some x_i for each $i \leq p$ if and only if $\{x^k\}$ converges to $x \equiv (x_1, \dots, x_p)$. See Theorem 3.7.3. ■

Also important is the following theorem.

Theorem 15.1.4 The set of terms in a Cauchy sequence in \mathbb{R}^p is bounded in the sense that for all n , $|x_n| < M$ for some $M < \infty$.

Proof: Let $\varepsilon = 1$ in the definition of a Cauchy sequence and let $n > n_1$. Then from the definition, $|x_n - x_{n_1}| < 1$. It follows that for all $n > n_1$, $|x_n| < 1 + |x_{n_1}|$. Therefore, for all n , $|x_n| \leq 1 + |x_{n_1}| + \sum_{k=1}^{n_1} |x_k|$ ■

Note that a sequence in \mathbb{R}^p is bounded if and only if the k^{th} components are bounded, this by 15.1.

15.2 Open and Closed Sets

Open sets are those sets S such that if $x \in S$, then so is y whenever y is sufficiently close to x . Closed sets are those sets S such that if $x_n \rightarrow x$ and each $x_n \in S$, then also $x \in S$. What follows is just a more precise statement of this.

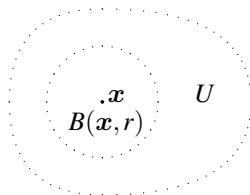
Eventually, one must consider functions which are defined on subsets of \mathbb{R}^p and their properties. The next definition will end up being quite important. It describes a type of subset of \mathbb{R}^p with the property that if x is in this set, then so is y whenever y is close enough to x .

Definition 15.2.1 Recall for $x, y \in \mathbb{R}^p$, $|x - y| = \left(\sum_{i=1}^p |x_i - y_i|^2\right)^{1/2}$. Also let $B(x, r) \equiv \{y \in \mathbb{R}^p : |x - y| < r\}$. Let $U \subseteq \mathbb{R}^p$. U is an **open set** if whenever $x \in U$, there exists $r > 0$ such that $B(x, r) \subseteq U$. More generally, if U is any subset of \mathbb{R}^p , $x \in U$ is an **interior point** of U if there exists $r > 0$ such that $x \in B(x, r) \subseteq U$. In other words U is an open set exactly when every point of U is an interior point of U .

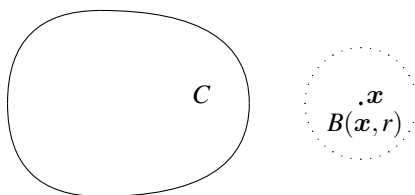
If there is something called an open set, surely there should be something called a closed set and here is the definition of one.

Definition 15.2.2 A subset, C , of \mathbb{R}^p is called a **closed set** if $\mathbb{R}^p \setminus C$ is an open set. The symbol $\mathbb{R}^p \setminus C$ denotes everything in \mathbb{R}^p which is not in C . It is also called the **complement** of C . The symbol S^C is a short way of writing $\mathbb{R}^p \setminus S$.

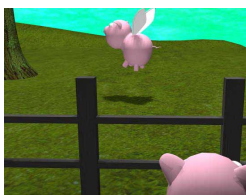
To illustrate this definition, consider the following picture.



You see in this picture how the edges are dotted. This is because an open set, can not include the edges or the set would fail to be open. For example, consider what would happen if you picked a point out on the edge of U in the above picture. Every open ball centered at that point would have in it some points which are outside U . Therefore, such a point would violate the above definition. You also see the edges of $B(x, r)$ dotted suggesting that $B(x, r)$ ought to be an open set. This is intuitively clear but does require a proof. This will be done in the next theorem and will give examples of open sets. Also, you can see that if x is close to the edge of U , you might have to take r to be very small. open sets do not have their skins while closed sets do. Here is a picture of a closed set, C .



Note that $x \notin C$ and since $\mathbb{R}^p \setminus C$ is open, there exists a ball, $B(x, r)$ contained entirely in $\mathbb{R}^p \setminus C$. If you look at $\mathbb{R}^p \setminus C$, what would be its skin? It can't be in $\mathbb{R}^p \setminus C$ and so it must be in C . This is a rough heuristic explanation of what is going on with these definitions. Also note that \mathbb{R}^p and \emptyset are both open and closed. Here is why. If $x \in \emptyset$, then there must be a ball centered at x which is also contained in \emptyset . This must be considered to be true because there is nothing in \emptyset so there can be no example to show it false¹. Therefore, from the definition, it follows \emptyset is open. It is also closed because if $x \notin \emptyset$, then $B(x, 1)$ is also contained in $\mathbb{R}^p \setminus \emptyset = \mathbb{R}^p$. Therefore, \emptyset is both open and closed. From this, it follows \mathbb{R}^p is also both open and closed.



Theorem 15.2.3 *Let $x \in \mathbb{R}^p$ and let $r \geq 0$. Then $B(x, r)$ is an open set. Also, $D(x, r) \equiv \{y \in \mathbb{R}^p : |y - x| \leq r\}$ is a closed set. In particular, every closed interval in \mathbb{R} is a closed set.*

Proof: Suppose $y \in B(x, r)$. It is necessary to show there exists $r_1 > 0$ such that $B(y, r_1) \subseteq B(x, r)$. Define $r_1 \equiv r - |x - y|$. Then if $|z - y| < r_1$, it follows from the above triangle inequality that

$$\begin{aligned} |z - x| &= |z - y + y - x| \leq |z - y| + |y - x| \\ &< r_1 + |y - x| = r - |x - y| + |y - x| = r. \end{aligned}$$

¹To a mathematician, the statement: Whenever a pig is born with wings it can fly must be taken as true. We do not consider biological or aerodynamic considerations in such statements. There is no such thing as a winged pig and therefore, all winged pigs must be superb flyers since there can be no example of one which is not. On the other hand we would also consider the statement: Whenever a pig is born with wings it cannot possibly fly, as equally true. The point is, you can say anything you want about the elements of the empty set and no one can gainsay your statement. Therefore, such statements are considered as true by default. You may say this is a very strange way of thinking about truth and ultimately this is because mathematics is not about truth. It is more about consistency and logic.

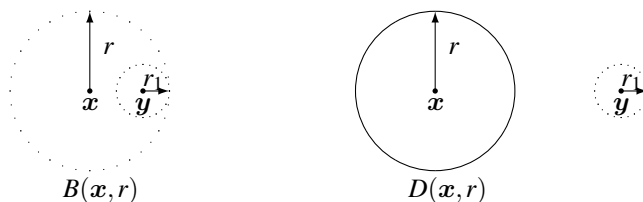
Note that if $r = 0$ then $B(x, r) = \emptyset$, the empty set. This is because if $y \in \mathbb{R}^p$, $|x - y| \geq 0$ and so $y \notin B(x, 0)$. Since \emptyset has no points in it, it must be open because every point in it, (There are none.) satisfies the desired property of being an interior point.

Now suppose $y \notin D(x, r)$. Then $|x - y| > r$ and defining $\delta \equiv |x - y| - r$, it follows that if $z \in B(y, \delta)$, then by the triangle inequality,

$$\begin{aligned} |x - z| &\geq |x - y| - |y - z| > |x - y| - \delta \\ &= |x - y| - (|x - y| - r) = r \end{aligned}$$

and this shows that $B(y, \delta) \subseteq \mathbb{R}^p \setminus D(x, r)$. Since y was an arbitrary point in $\mathbb{R}^p \setminus D(x, r)$, it follows $\mathbb{R}^p \setminus D(x, r)$ is an open set which shows, from the definition, that $D(x, r)$ is a closed set as claimed. Now $[a, b] = D\left(\frac{a+b}{2}, \frac{b-a}{2}\right)$. ■

A picture which is descriptive of the conclusion of the above theorem which also implies the manner of proof is the following.



The next theorem includes the main ideas for a set to be closed. It says that closed is to retain all limits of sequences which are contained in A .

Theorem 15.2.4 *A nonempty set A is closed if and only if whenever $x_k \in A$ and $\lim_{k \rightarrow \infty} x_k = x$, it follows that $x \in A$. In other words, the set is closed if and only if every convergent sequence of points of A converges to a point of A .*

Proof: Suppose A is closed and suppose $\lim_{k \rightarrow \infty} x_k = x$. Does it follow that $x \in A$? If not, then since A is closed, its complement is open and so there is a ball $B(x, r)$ contained in A^C . However, this contradicts the assertion that x is the limit of the sequence. Indeed, x_k must be in $B(x, r)$ for all k sufficiently large.

Conversely, suppose A retains all limits of convergent sequences. Is A closed? In other words, is its complement A^C open? Suppose $x \in A^C$. Is $B(x, r) \subseteq A^C$ for small enough positive r ? If not, then $B(x, \frac{1}{k})$ contains a point of A called x_k for each $k = 1, 2, \dots$. Thus x is a limit of the sequence $\{x_k\}$ and so $x \in A$ after all. Hence A^C must indeed be open and so, by definition, A is closed. ■

15.3 Cartesian Products

Recall \mathbb{R}^2 consists of ordered pairs (x, y) such that $x \in \mathbb{R}$ and $y \in \mathbb{R}$. \mathbb{R}^2 is also written as $\mathbb{R} \times \mathbb{R}$. In general, the following definition holds.

Definition 15.3.1 *The Cartesian product of two sets $A \times B$, means*

$$\{(a, b) : a \in A, b \in B\}.$$

If you have n sets A_1, A_2, \dots, A_n

$$\prod_{i=1}^n A_i = \{(x_1, x_2, \dots, x_n) : \text{each } x_i \in A_i\}.$$

Now suppose $A \subseteq \mathbb{R}^m$ and $B \subseteq \mathbb{R}^p$. Then if $(\mathbf{x}, \mathbf{y}) \in A \times B$, $\mathbf{x} = (x_1, \dots, x_m)$, and $\mathbf{y} = (y_1, \dots, y_p)$, the following identification will be made.

$$(\mathbf{x}, \mathbf{y}) = (x_1, \dots, x_m, y_1, \dots, y_p) \in \mathbb{R}^{p+m}.$$

Similarly, starting with something in \mathbb{R}^{p+m} , you can write it in the form (\mathbf{x}, \mathbf{y}) where $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^p$. The following theorem has to do with the Cartesian product of two closed sets or two open sets. Also here is an important definition.

Definition 15.3.2 A set, $A \subseteq \mathbb{R}^p$ is said to be **bounded** if there exist finite intervals, $[a_i, b_i]$ such that $A \subseteq \prod_{i=1}^p [a_i, b_i]$.

Theorem 15.3.3 Let U be an open set in \mathbb{R}^m and let V be an open set in \mathbb{R}^p . Then $U \times V$ is an open set in \mathbb{R}^{p+m} . If C is a closed set in \mathbb{R}^m and H is a closed set in \mathbb{R}^p , then $C \times H$ is a closed set in \mathbb{R}^{p+m} . If C and H are bounded, then so is $C \times H$.

Proof: Let $(\mathbf{x}, \mathbf{y}) \in U \times V$. Since U is open, there exists $r_1 > 0$ such that $B(\mathbf{x}, r_1) \subseteq U$. Similarly, there exists $r_2 > 0$ such that $B(\mathbf{y}, r_2) \subseteq V$. Now

$$B((\mathbf{x}, \mathbf{y}), \delta) \equiv \left\{ (\mathbf{s}, \mathbf{t}) \in \mathbb{R}^{p+m} : \sum_{k=1}^m |x_k - s_k|^2 + \sum_{j=1}^p |y_j - t_j|^2 < \delta^2 \right\}$$

Therefore, if $\delta \equiv \min(r_1, r_2)$ and $(\mathbf{s}, \mathbf{t}) \in B((\mathbf{x}, \mathbf{y}), \delta)$, then it follows that $\mathbf{s} \in B(\mathbf{x}, r_1) \subseteq U$ and that $\mathbf{t} \in B(\mathbf{y}, r_2) \subseteq V$ which shows that $B((\mathbf{x}, \mathbf{y}), \delta) \subseteq U \times V$. Hence $U \times V$ is open as claimed.

Next suppose $(\mathbf{x}, \mathbf{y}) \notin C \times H$. It is necessary to show there exists $\delta > 0$ such that $B((\mathbf{x}, \mathbf{y}), \delta) \subseteq \mathbb{R}^{p+m} \setminus (C \times H)$. Either $\mathbf{x} \notin C$ or $\mathbf{y} \notin H$ since otherwise (\mathbf{x}, \mathbf{y}) would be a point of $C \times H$. Suppose therefore, that $\mathbf{x} \notin C$. Since C is closed, there exists $r > 0$ such that $B(\mathbf{x}, r) \subseteq \mathbb{R}^m \setminus C$. Consider $B((\mathbf{x}, \mathbf{y}), r)$. If $(\mathbf{s}, \mathbf{t}) \in B((\mathbf{x}, \mathbf{y}), r)$, it follows that $\mathbf{s} \in B(\mathbf{x}, r)$ which is contained in $\mathbb{R}^m \setminus C$. Therefore, $B((\mathbf{x}, \mathbf{y}), r) \subseteq \mathbb{R}^{p+m} \setminus (C \times H)$ showing $C \times H$ is closed. A similar argument holds if $\mathbf{y} \notin H$.

If C is bounded, there exist $[a_i, b_i]$ such that $C \subseteq \prod_{i=1}^m [a_i, b_i]$ and if H is bounded, $H \subseteq \prod_{i=m+1}^{m+p} [a_i, b_i]$ for intervals $[a_{m+1}, b_{m+1}], \dots, [a_{m+p}, b_{m+p}]$. Therefore, $C \times H \subseteq \prod_{i=1}^{m+p} [a_i, b_i]$. ■

15.4 Sequential Compactness

The concept of sequential compactness is also the same as before. I will show here that, as before, the sequentially compact sets are closed and bounded.

Definition 15.4.1 A set K in \mathbb{R}^p is sequentially compact if every sequence in K has a subsequence which converges to a point in K .

Theorem 15.4.2 *Let K be a nonempty subset of \mathbb{R}^p . Then K is sequentially compact if and only if it is closed and bounded.*

Proof: Suppose first that K is closed and bounded. Then by definition, $K \subseteq \prod_{i=1}^p [a_i, b_i]$ for a suitable product of closed and bounded intervals. Let the sequence be $\{\mathbf{x}^k\}_{k=1}^\infty$, $\mathbf{x}^k = (x_1^k, \dots, x_p^k)$. Then it follows from the definition of the Cartesian product that for each i , $x_i^k \in [a_i, b_i]$ for all k . Then $\{x_1^k\}_{k=1}^\infty$ has a convergent subsequence, denoted by $x_1^{k_1}$ such that $\lim_{k_1 \rightarrow \infty} x_1^{k_1} = x_1$. Now $\{x_2^{k_1}\}_{k_1=1}^\infty$ has a convergent subsequence denoted as $\{x_2^{k_2}\}_{k_2=1}^\infty$ converging to $x_2 \in [a_2, b_2]$. Recall that if a sequence of real numbers converges, then so does every subsequence. It follows that $\lim_{k_2 \rightarrow \infty} x_1^{k_2} = x_1$. Continue taking subsequences such that $\lim_{k_r \rightarrow \infty} x_j^{k_r} = x_j \in [a_j, b_j]$ for each $j \leq r$. Therefore, $\lim_{k_p \rightarrow \infty} x_i^{k_p} = x_i$ for each $i \leq p$ and this shows that $\lim_{k_p \rightarrow \infty} \mathbf{x}^{k_p} = \mathbf{x}$ where $\mathbf{x} = (x_1, \dots, x_p) \in \prod_{i=1}^p [a_i, b_i]$. However, K is closed and so $\mathbf{x} \in K$. This shows that a closed and bounded nonempty set is sequentially compact.

Conversely, suppose a set K is sequentially compact. Then the set must be bounded since otherwise one could obtain a sequence of points $\{\mathbf{x}^n\}_{n=1}^\infty$ with $|\mathbf{x}^n| > n$. Thus every subsequence is unbounded so no subsequence can be a Cauchy sequence and so no subsequence can converge. If the set is not closed, then by Theorem 15.2.4 above, there would be a point $\mathbf{x} \notin K$ and a sequence of points of K $\{\mathbf{x}_k\}_{k=1}^\infty$ which converges to \mathbf{x} . But now this sequence must have a convergent subsequence converging to a point of K . This is impossible because all subsequences must converge to \mathbf{x} which is not in K . Therefore, K must also be closed. ■

15.5 Vector Valued Functions

Vector valued functions have values in \mathbb{R}^p where p is an integer at least as large as 1. Here are some examples.

Example 15.5.1 *A rocket is launched from the rotating earth. You could define a function having values in \mathbb{R}^3 as $(r(t), \theta(t), \phi(t))$ where $r(t)$ is the distance of the center of mass of the rocket from the center of the earth, $\theta(t)$ is the longitude, and $\phi(t)$ is the latitude of the rocket.*

Example 15.5.2 *Let $\mathbf{f}(x, y) = (\sin xy, y^3 + x, x^4)$. Then \mathbf{f} is a function defined on \mathbb{R}^2 which has values in \mathbb{R}^3 . For example, $\mathbf{f}(1, 2) = (\sin 2, 9, 16)$.*

As usual, $D(\mathbf{f})$ denotes the domain of the function \mathbf{f} which is written in bold face because it will possibly have values in \mathbb{R}^p . When $D(\mathbf{f})$ is not specified, it will be understood that the domain of \mathbf{f} consists of those things for which \mathbf{f} makes sense.

Example 15.5.3 *Let $\mathbf{f}(x, y, z) = \left(\frac{x+y}{z}, \sqrt{1-x^2}, y\right)$. Then $D(\mathbf{f})$ would consist of the set of all (x, y, z) such that $|x| \leq 1$ and $z \neq 0$.*

There are many ways to make new functions from old ones.

Definition 15.5.4 Let \mathbf{f}, \mathbf{g} be functions with values in \mathbb{R}^p . Let a, b be points of \mathbb{R} (scalars). Then $a\mathbf{f} + b\mathbf{g}$ is the name of a function whose domain is $D(\mathbf{f}) \cap D(\mathbf{g})$ which is defined as $(a\mathbf{f} + b\mathbf{g})(x) = a\mathbf{f}(x) + b\mathbf{g}(x)$. Also, $\mathbf{f} \cdot \mathbf{g}$ or (\mathbf{f}, \mathbf{g}) is the name of a function whose domain is $D(\mathbf{f}) \cap D(\mathbf{g})$ which is defined as $(\mathbf{f}, \mathbf{g})(x) \equiv \mathbf{f} \cdot \mathbf{g}(x) \equiv \mathbf{f}(x) \cdot \mathbf{g}(x)$. If \mathbf{f} and \mathbf{g} have values in \mathbb{R}^3 , define a new function $\mathbf{f} \times \mathbf{g}$ by $\mathbf{f} \times \mathbf{g}(t) \equiv \mathbf{f}(t) \times \mathbf{g}(t)$. If $\mathbf{f} : D(\mathbf{f}) \rightarrow X$ and $\mathbf{g} : X \rightarrow Y$, then $\mathbf{g} \circ \mathbf{f}$ is the name of a function whose domain is $\{x \in D(\mathbf{f}) : \mathbf{f}(x) \in D(\mathbf{g})\}$ which is defined as $\mathbf{g} \circ \mathbf{f}(x) \equiv \mathbf{g}(\mathbf{f}(x))$. This is called the composition of the two functions.

You should note that $\mathbf{f}(x)$ is not a function. It is the value of the function at the point x . The name of the function is \mathbf{f} . Nevertheless, people often write $\mathbf{f}(x)$ to denote a function and it does not cause too many problems in beginning courses. When this is done, the variable, x should be considered as a generic variable which is allowed to be anything in $D(\mathbf{f})$. I will use this slightly sloppy abuse of notation whenever convenient.

Example 15.5.5 Let $\mathbf{f}(t) \equiv (t, 1+t, 2)$ and $\mathbf{g}(t) \equiv (t^2, t, t)$. Then $\mathbf{f} \cdot \mathbf{g}$ is the name of the function satisfying $\mathbf{f} \cdot \mathbf{g}(t) = \mathbf{f}(t) \cdot \mathbf{g}(t) = t^3 + t + t^2 + 2t = t^3 + t^2 + 3t$.

Note that in this case it was assumed the domains of the functions consisted of all of \mathbb{R} because this was the set on which the two both made sense. Also note that \mathbf{f} and \mathbf{g} map \mathbb{R} into \mathbb{R}^3 but $\mathbf{f} \cdot \mathbf{g}$ maps \mathbb{R} into \mathbb{R} .

Example 15.5.6 Suppose $\mathbf{f}(t) = (2t, 1+t^2)$ and $\mathbf{g} : \mathbb{R}^2 \rightarrow \mathbb{R}$ is given by $\mathbf{g}(x, y) \equiv x + y$. Then $\mathbf{g} \circ \mathbf{f} : \mathbb{R} \rightarrow \mathbb{R}$ and $\mathbf{g} \circ \mathbf{f}(t) = \mathbf{g}(\mathbf{f}(t)) = \mathbf{g}(2t, 1+t^2) = 1 + 2t + t^2$.

15.6 Continuous Functions

What was done in one variable calculus for scalar functions is generalized here to include the case of a vector valued function of possibly many variables. This part of the book is on functions of a single variable. However, it is no harder to consider the limit and continuity in terms of a function of many variables and it seems a good idea to go ahead and do it.

Definition 15.6.1 A function $\mathbf{f} : D(\mathbf{f}) \subseteq \mathbb{R}^p \rightarrow \mathbb{R}^q$ is continuous at $\mathbf{x} \in D(\mathbf{f})$ if for each $\varepsilon > 0$ there exists $\delta > 0$ such that whenever $\mathbf{y} \in D(\mathbf{f})$ and $|\mathbf{y} - \mathbf{x}| < \delta$ it follows that $|\mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{x})| < \varepsilon$. \mathbf{f} is continuous if it is continuous at every point of $D(\mathbf{f})$.

Note the total similarity to the scalar valued case. Also one obtains a similar description in terms of convergent sequences.

Definition 15.6.2 $\lim_{n \rightarrow \infty} \mathbf{x}_n = \mathbf{x}$ means: For every $\varepsilon > 0$ there is n_ε such that if $n \geq n_\varepsilon$, then $|\mathbf{x}_n - \mathbf{x}| < \varepsilon$. From the description of $|\cdot|$ given earlier, this says the same as $\lim_{n \rightarrow \infty} x_n^i = x^i$ for each $i = 1, 2, \dots, p$ where $\mathbf{x}_n \equiv (x_n^1, \dots, x_n^p)$ and $\mathbf{x} \equiv (x^1, \dots, x^p)$.

A repeat of the earlier theorem for functions of one variable yields the following equivalent description of continuity. All you have to do is make things bold face and repeat the earlier argument.

Proposition 15.6.3 $\mathbf{f} : D(\mathbf{f}) \rightarrow \mathbb{R}^q$ is continuous at $\mathbf{x} \in D(\mathbf{f})$ means that whenever $\mathbf{x}_n \rightarrow \mathbf{x}$ with $\mathbf{x}_n \in D(\mathbf{f})$, it follows that $\mathbf{f}(\mathbf{x}_n) \rightarrow \mathbf{f}(\mathbf{x})$.

Proof: \Rightarrow Suppose f is continuous at x and $x_n \rightarrow x$. Given $\varepsilon > 0$, let δ correspond to ε in the definition of continuity. Then for all n large enough, $|x_n - x| < \delta$ and so for all n large enough, $|f(x) - f(x_n)| < \varepsilon$. Thus $f(x_n) \rightarrow f(x)$ by definition.

\Leftarrow Suppose the condition of taking convergent sequences to convergent sequences at x . If f is not continuous, then there exists $\varepsilon > 0$ such that for any $\delta > 0$ there will be a x_δ such that although $|x_\delta - x| < \delta$, $|f(x_\delta) - f(x)| \geq \varepsilon$. Now let x_n equal the exceptional point corresponding to $\delta = 1/n$, $n \in \mathbb{N}$. Then $x_n \rightarrow x$ but $f(x_n)$ fails to converge to $f(x)$ which is a contradiction. Thus, it can't happen that the function fails to be continuous at x . ■

In the following important proposition, $\|\cdot\|$ will be a norm on \mathbb{R}^p . It could be the usual one $|\cdot|$ being the square root of the sum of the squares or it could be $\|\cdot\|_\infty$ given by $\|x\|_\infty = \max\{|x_i| : i \leq p\}$ or any other norm. The notion is completely general. However, go ahead and restrict to $|\cdot|$ if this is causing confusion.

Proposition 15.6.4 *Let S be a nonempty set and let*

$$\text{dist}(x, S) \equiv \inf\{\|x - s\| : s \in S\}$$

Then $x \rightarrow \text{dist}(x, S)$ is continuous. In fact, $|\text{dist}(x, S) - \text{dist}(y, S)| \leq \|x - y\|$.

Proof: Say $\text{dist}(x, S) - \text{dist}(y, S) > 0$. Then pick $s \in S$ such that $\|y - s\| - \varepsilon < \text{dist}(y, S)$. Then

$$\begin{aligned} |\text{dist}(x, S) - \text{dist}(y, S)| &= \text{dist}(x, S) - \text{dist}(y, S) \leq \text{dist}(x, S) - \|y - s\| + \varepsilon \\ &\leq \|x - s\| - \|y - s\| + \varepsilon \leq \|x - y\| + \|y - s\| - \|y - s\| + \varepsilon \\ &= \|x - y\| + \varepsilon \end{aligned}$$

Since ε is arbitrary, this shows the inequality and proves continuity. ■

15.7 Sufficient Conditions for Continuity

The next theorem is a fundamental result which allows less worry about the $\varepsilon \delta$ definition of continuity.

Theorem 15.7.1 *The following assertions are valid.*

1. *The function $af + bg$ is continuous at x whenever f, g are continuous at $x \in D(f) \cap D(g)$ and $a, b \in \mathbb{R}$.*
2. *If f is continuous at x , $f(x) \in D(g) \subseteq \mathbb{R}^p$, and g is continuous at $f(x)$, then $g \circ f$ is continuous at x .*
3. *If $f = (f_1, \dots, f_q) : D(f) \rightarrow \mathbb{R}^q$, then f is continuous if and only if each f_k is a continuous real valued function.*
4. *The function $f : \mathbb{R}^p \rightarrow \mathbb{R}$, given by $f(x) = |x|$ is continuous.*
5. *The map $\pi_k(x) \equiv x_k$ is continuous.*
6. *Every function $x \rightarrow x_1^{\alpha_1} x_2^{\alpha_2} \dots x_p^{\alpha_p}$ for α_i an integer is continuous.*

7. If f, g are each continuous at x , then $f \cdot g$ is also continuous at x .

This is proved just like the corresponding theorem for functions of a single variable. For example the first claim says that $(af + bg)(y)$ is close to $(af + bg)(x)$ when y is close to x provided the same can be said about f and g . For the second claim, if y is close to x , $f(x)$ is close to $f(y)$ and so by continuity of g at $f(x)$, $g(f(y))$ is close to $g(f(x))$. To see the third claim is likely, note that closeness in \mathbb{R}^p is the same as closeness in each coordinate. The fourth claim is immediate from the triangle inequality. Alternatively, use Proposition 15.6.3 to reduce to notions of convergent sequences and then Definition 15.6.2 to reduce completely to one variable considerations and apply earlier theorems on limits and continuity.

For functions defined on \mathbb{R}^p , there is a notion of polynomial just as there is for functions defined on \mathbb{R} .

Definition 15.7.2 Let α be an p dimensional multi-index. This means

$$\alpha = (\alpha_1, \dots, \alpha_p)$$

where each α_i is a natural number or zero. Also, let

$$|\alpha| \equiv \sum_{i=1}^p |\alpha_i|$$

The symbol x^α means

$$x^\alpha \equiv x_1^{\alpha_1} x_2^{\alpha_2} \dots x_p^{\alpha_p}.$$

An p dimensional polynomial of degree m is a function of the form

$$p(x) = \sum_{|\alpha| \leq m} d_\alpha x^\alpha.$$

where the d_α are real numbers.

The above Theorem 15.7.1 implies that polynomials are all continuous. Also, rational functions, being quotients of polynomials are also continuous at every point where the denominator is not zero. This follows from the theorems on sequences presented earlier and the above.

15.8 Limits of a Function of Many Variables

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 x , which are limit points of $D(f)$ and this concept is defined next.

Definition 15.8.1 Let $A \subseteq \mathbb{R}^m$ be a set. A point x , is a limit point of A if $B(x, r)$ contains infinitely many points of A for every $r > 0$.

Definition 15.8.2 Let $f : D(f) \subseteq \mathbb{R}^p \rightarrow \mathbb{R}^q$ be a function and let x be a limit point of $D(f)$. Then

$$\lim_{y \rightarrow x} f(y) = L$$

if and only if the following condition holds. For all $\varepsilon > 0$ there exists $\delta > 0$ such that if

$$0 < |y - x| < \delta, \text{ and } y \in D(f)$$

then,

$$|L - f(y)| < \varepsilon.$$

Theorem 15.8.3 If $\lim_{y \rightarrow x} f(y) = L$ and $\lim_{y \rightarrow x} f(y) = L_1$, then $L = L_1$.

Proof: Let $\varepsilon > 0$ be given. There exists $\delta > 0$ such that if $0 < |y - x| < \delta$ and $y \in D(f)$, then

$$|f(y) - L| < \varepsilon, |f(y) - L_1| < \varepsilon.$$

Pick such a y . There exists one because x is a limit point of $D(f)$. Then

$$|L - L_1| \leq |L - f(y)| + |f(y) - L_1| < \varepsilon + \varepsilon = 2\varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, this shows $L = L_1$. ■

One can define what it means for $\lim_{y \rightarrow x} f(x) = \pm\infty$ just as for sequences.

Definition 15.8.4 If $f(x) \in \mathbb{R}$, $\lim_{y \rightarrow x} f(x) = \infty$ if for every number l , there exists $\delta > 0$ such that whenever $0 < |y - x| < \delta$ and $y \in D(f)$, then $f(x) > l$. $\lim_{y \rightarrow x} f(x) = -\infty$ if for every number l , there exists $\delta > 0$ such that whenever $0 < |y - x| < \delta$ and $y \in D(f)$, then $f(x) < l$.

As before, it is useful to reduce to a statement about sequences.

Proposition 15.8.5 Let x be a limit point of $D(f)$. Then $\lim_{y \rightarrow x} f(y) = L$ if and only if whenever $x_n \rightarrow x$ for each $x_n \neq x$, the x_n distinct points, it follows that $f(x_n) \rightarrow L$.

Proof: \Rightarrow Let $x_n \rightarrow x$ where no x_n equals x . Let $\varepsilon > 0$ be given. By assumption, $|f(y) - L| < \varepsilon$ whenever $0 < |y - x| < \delta$ for some δ . However, for all n large enough, $0 < |x_n - x| < \delta$ and so $|f(x_n) - L| < \varepsilon$. Hence $f(x_n) \rightarrow L$.

\Leftarrow Suppose the condition on the sequences holds. If the condition for the limit does not hold, then there exists $\varepsilon > 0$ such that no matter how small δ , there will be $0 < |y - x| < \delta$, $y \in D(f)$, and yet $|f(y) - L| \geq \varepsilon$. Now let $\delta_1 = 1$. There exists $x_1 \neq x$ with $x_1 \in B(x, \delta_1) \cap D(f)$ and $|f(x_1) - L| \geq \varepsilon$. Let $\delta_2 \equiv \min(\frac{1}{2}, \frac{1}{2}|x - x_1|)$. Now pick $x_2 \in B(x, \delta_2)$, $x_2 \neq x$ such that $|f(x_2) - L| \geq \varepsilon$. Let $\delta_3 \equiv \min(\frac{1}{2^3}, \frac{1}{2}|x - x_1|, \frac{1}{2}|x - x_2|)$ and pick $x_3 \in B(x, \delta_3)$ with $|f(x_3) - L| \geq \varepsilon$, $x_3 \neq x$. Continue this way to generate a sequence of distinct points $\{x_n\}$, none equal to x which converges to x . Then $L = \lim_{n \rightarrow \infty} f(x_n)$ because of the condition on limits of the sequence so eventually

$$|L - f(x_n)| < \varepsilon,$$

contrary to the construction of the x_n . ■

The following theorem is just like the one variable version calculus.

Theorem 15.8.6 Suppose $f : D(f) \rightarrow \mathbb{R}^q$. Then for x a limit point of $D(f)$,

$$\lim_{y \rightarrow x} f(y) = L \tag{15.2}$$

if and only if

$$\lim_{\mathbf{y} \rightarrow \mathbf{x}} f_k(\mathbf{y}) = L_k \quad (15.3)$$

where $\mathbf{f}(\mathbf{y}) \equiv (f_1(\mathbf{y}), \dots, f_p(\mathbf{y}))$ and $\mathbf{L} \equiv (L_1, \dots, L_p)$. Suppose

$$\lim_{\mathbf{y} \rightarrow \mathbf{x}} \mathbf{f}(\mathbf{y}) = \mathbf{L}, \quad \lim_{\mathbf{y} \rightarrow \mathbf{x}} \mathbf{g}(\mathbf{y}) = \mathbf{K}$$

where $\mathbf{K}, \mathbf{L} \in \mathbb{R}^q$. Then if $a, b \in \mathbb{R}$,

$$\lim_{\mathbf{y} \rightarrow \mathbf{x}} (a\mathbf{f}(\mathbf{y}) + b\mathbf{g}(\mathbf{y})) = a\mathbf{L} + b\mathbf{K}, \quad (15.4)$$

$$\lim_{\mathbf{y} \rightarrow \mathbf{x}} \mathbf{f}(\mathbf{y}) \cdot \mathbf{g}(\mathbf{y}) = \mathbf{L} \cdot \mathbf{K} \quad (15.5)$$

In the case where $q = 3$ and $\lim_{\mathbf{y} \rightarrow \mathbf{x}} \mathbf{f}(\mathbf{y}) = \mathbf{L}$ and $\lim_{\mathbf{y} \rightarrow \mathbf{x}} \mathbf{g}(\mathbf{y}) = \mathbf{K}$, then

$$\lim_{\mathbf{y} \rightarrow \mathbf{x}} \mathbf{f}(\mathbf{y}) \times \mathbf{g}(\mathbf{y}) = \mathbf{L} \times \mathbf{K}. \quad (15.6)$$

If g is scalar valued with $\lim_{\mathbf{y} \rightarrow \mathbf{x}} g(\mathbf{y}) = K \neq 0$,

$$\lim_{\mathbf{y} \rightarrow \mathbf{x}} \mathbf{f}(\mathbf{y}) g(\mathbf{y}) = \mathbf{L}K. \quad (15.7)$$

Also, if h is a continuous function defined near \mathbf{L} , then

$$\lim_{\mathbf{y} \rightarrow \mathbf{x}} h \circ \mathbf{f}(\mathbf{y}) = h(\mathbf{L}). \quad (15.8)$$

Suppose $\lim_{\mathbf{y} \rightarrow \mathbf{x}} \mathbf{f}(\mathbf{y}) = \mathbf{L}$. If $|\mathbf{f}(\mathbf{y}) - \mathbf{b}| \leq r$ for all \mathbf{y} sufficiently close to \mathbf{x} , then $|\mathbf{L} - \mathbf{b}| \leq r$ also.

Proof: All of these claims follow from consideration of components and the properties of limits of sequences and Proposition 15.8.5. As an example, consider the last claim. Let $\mathbf{x}_n \rightarrow \mathbf{x}$ where the \mathbf{x}_n are distinct and none equal to \mathbf{x} . Then $\mathbf{f}(\mathbf{x}_n) \rightarrow \mathbf{L}$ and so by continuity of h at \mathbf{L} , $h(\mathbf{f}(\mathbf{x}_n)) \rightarrow h(\mathbf{L})$.

The relation between continuity and limits is as follows.

Theorem 15.8.7 For $\mathbf{f} : D(\mathbf{f}) \rightarrow \mathbb{R}^q$ and $\mathbf{x} \in D(\mathbf{f})$ a limit point of $D(\mathbf{f})$, \mathbf{f} is continuous at \mathbf{x} if and only if

$$\lim_{\mathbf{y} \rightarrow \mathbf{x}} \mathbf{f}(\mathbf{y}) = \mathbf{f}(\mathbf{x}).$$

Proof: First suppose \mathbf{f} is continuous at \mathbf{x} a limit point of $D(\mathbf{f})$. Then for every $\varepsilon > 0$ there exists $\delta > 0$ such that if $|\mathbf{y} - \mathbf{x}| < \delta$ and $\mathbf{y} \in D(\mathbf{f})$, then $|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| < \varepsilon$. In particular, this holds if $0 < |\mathbf{x} - \mathbf{y}| < \delta$ and this is just the definition of the limit. Hence $\mathbf{f}(\mathbf{x}) = \lim_{\mathbf{y} \rightarrow \mathbf{x}} \mathbf{f}(\mathbf{y})$.

Next suppose \mathbf{x} is a limit point of $D(\mathbf{f})$ and $\lim_{\mathbf{y} \rightarrow \mathbf{x}} \mathbf{f}(\mathbf{y}) = \mathbf{f}(\mathbf{x})$. This means that if $\varepsilon > 0$ there exists $\delta > 0$ such that for $0 < |\mathbf{x} - \mathbf{y}| < \delta$ and $\mathbf{y} \in D(\mathbf{f})$, it follows $|\mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{x})| < \varepsilon$. However, if $\mathbf{y} = \mathbf{x}$, then $|\mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{x})| = |\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x})| = 0$ and so whenever $\mathbf{y} \in D(\mathbf{f})$ and $|\mathbf{x} - \mathbf{y}| < \delta$, it follows $|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| < \varepsilon$, showing \mathbf{f} is continuous at \mathbf{x} . ■

Example 15.8.8 Find $\lim_{(x,y) \rightarrow (3,1)} \left(\frac{x^2-9}{x-3}, y \right)$.

It is clear that $\lim_{(x,y) \rightarrow (3,1)} \frac{x^2-9}{x-3} = 6$ and $\lim_{(x,y) \rightarrow (3,1)} y = 1$. Therefore, this limit equals $(6, 1)$.

Example 15.8.9 Find $\lim_{(x,y) \rightarrow (0,0)} \frac{xy}{x^2+y^2}$.

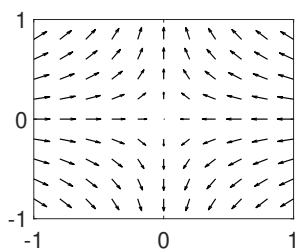
First of all, observe the domain of the function is $\mathbb{R}^2 \setminus \{(0,0)\}$, every point in \mathbb{R}^2 except the origin. Therefore, $(0,0)$ is a limit point of the domain of the function so it might make sense to take a limit. However, just as in the case of a function of one variable, 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.

15.9 Vector Fields

Some people find it useful to try and draw pictures to illustrate a vector valued function. This can be a very useful idea in the case where the function takes points in $D \subseteq \mathbb{R}^2$ and delivers a vector in \mathbb{R}^2 . For many points $(x,y) \in D$, you draw an arrow of the appropriate length and direction with its tail at (x,y) . The picture of all these arrows can give you an understanding of what is happening. For example if the vector valued function gives the velocity of a fluid at the point (x,y) , the picture of these arrows can give an idea of the motion of the fluid. When they are long the fluid is moving fast, when they are short, the fluid is moving slowly. The direction of these arrows is an indication of the direction of motion. The only sensible way to produce such a picture is with a computer. Otherwise, it becomes a worthless exercise in busy work. Furthermore, it is of limited usefulness in three dimensions because in three dimensions such pictures are too cluttered to convey much insight.

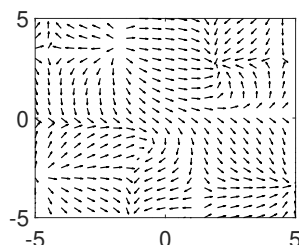
Example 15.9.1 Draw a picture of the vector field $(-x,y)$ which gives the velocity of a fluid flowing in two dimensions.



You can see how the arrows indicate the motion of this fluid.

Here is another such example. This one is much more complicated.

Example 15.9.2 Draw a picture of the vector field $(y\cos(x) + 1, x\sin(y) - 1)$ for the velocity of a fluid flowing in two dimensions.



Note how they reveal both the direction and the magnitude of the vectors. However, if you try to draw these by hand, you will mainly waste time.

15.10 MATLAB and Vector Fields

As mentioned, you should use a computer algebra system to graph vector fields. Here is an example of how to do this in MATLAB. Remember that to go to a new line, you press shift enter and to get it to do something, you press enter.

```
>>[a,b]=meshgrid(-1.5:.2:1.5,-1.5:.2:1.5);
u=b+a.^2.*b; v=-(b+2*a)+a.^3; r=(u.*u+v.*v+.1).^(1/2);
figure
quiver(a,b,u./r,v./r,'autoscalefactor',.5)
```

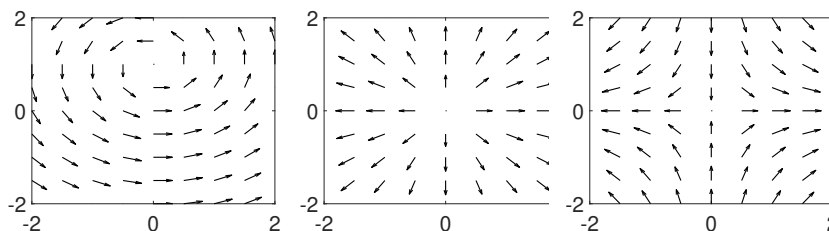
The .2 in the top line tells how close the vectors should be. This one graphs the vector field $(y + x^2y, -(y + 2x) + x^3)$. If you leave off the division by r you will see the relative size of the vectors. I have divided by r to expose only the direction. I have found that if I don't do this, the arrows get so small, I can't see them well. Of course, this is useful because it indicates a point of stagnation.

15.11 Exercises

- Here are some vector valued functions.

$$\mathbf{f}(x,y) = (x,y), \mathbf{g}(x,y) = (-(y-1),x), \mathbf{h}(x,y) = (x,-y).$$

Now here are the graphs of some vector fields. Match the function with the vector field.



- Find $D(\mathbf{f})$ for $\mathbf{f}(x,y,z,w) = \left(\frac{xy}{zw}, \sqrt{6-x^2y^2}\right)$.

3. Find $D(\mathbf{f})$ for $\mathbf{f}(x, y, z) = \left(\frac{1}{1+x^2-y^2}, \sqrt{4-(x^2+y^2+z^2)} \right)$.

4. For $\mathbf{f}(x, y, z) = (x, y, xy)$, $\mathbf{h}(x, y, z) = (y^2, -x, z)$ and

$$\mathbf{g}(x, y, z) = \left(\frac{1}{x}, yz, x^2 - 1 \right)$$

, compute the following.

(a) $\mathbf{f} \times \mathbf{g}$

(d) $\mathbf{f} \times \mathbf{g} \cdot \mathbf{h}$

(b) $\mathbf{g} \times \mathbf{f}$

(e) $\mathbf{f} \times (\mathbf{g} \times \mathbf{h})$

(c) $\mathbf{f} \cdot \mathbf{g}$

(f) $(\mathbf{f} \times \mathbf{g}) \cdot (\mathbf{g} \times \mathbf{h})$

5. Let $\mathbf{f}(x, y, z) = (y, z, x)$ and $\mathbf{g}(x, y, z) = (x^2 + y, z, x)$. Find $\mathbf{g} \circ \mathbf{f}(x, y, z)$.

6. Let $\mathbf{f}(x, y, z) = (x, z, yz)$ and $\mathbf{g}(x, y, z) = (x, y, x^2 - 1)$. Find $\mathbf{g} \circ \mathbf{f}(x, y, z)$.

7. For $\mathbf{f}, \mathbf{g}, \mathbf{h}$ vector valued functions and k, l scalar valued functions, which of the following make sense?

(a) $\mathbf{f} \times \mathbf{g} \times \mathbf{h}$

(d) $(\mathbf{f} \times \mathbf{g}) \cdot \mathbf{h}$

(b) $(k \times \mathbf{g}) \times \mathbf{h}$

(e) $l \cdot \mathbf{g} \cdot k$

(c) $(\mathbf{f} \cdot \mathbf{g}) \times \mathbf{h}$

(f) $\mathbf{f} \times (\mathbf{g} + \mathbf{h})$

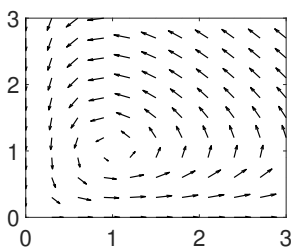
8. The Lotka Volterra system of differential equations, proposed in 1925 and 1926 by Lotka and Volterra respectively, is intended to model the interaction of predators and prey. An example of this situation is that of wolves and moose living on Isle Royal in the middle of Lake Superior. In these equations x is the number of prey and y is the number of predators. The equations are

$$x'(t) = x(t)(a - by(t)), \quad y'(t) = -y(t)(c - dx(t))$$

Written in terms of vectors,

$$(x', y') = (x(a - by), -y(c - dx))$$

The parameters a, b, c, d depend on the problem. The differential equations are saying that at a point (x, y) , the population vector (x, y) moves in the direction of $(x(a - by), -y(c - dx))$. Here is the graph of the vector field which determines the Lotka Volterra system in the case where all the parameters equal 1 which is graphed near the point $(1, 1)$. What conclusions seem to be true based on the graph of this vector field? What happens if you start with a population vector near the point $(1, 1)$? Remember these vectors in the plane determine the directions of motion of the population vector.



How did I know to graph the vector field near (1,1)?

15.12 Extreme Value Theorem, Uniform Continuity

Definition 15.12.1 A function \mathbf{f} having values in \mathbb{R}^p for $\mathbf{x} \in D$ is said to be bounded if the set of values of \mathbf{f} is a bounded set, meaning that if

$$\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}) \quad \cdots \quad f_p(\mathbf{x})),$$

then each $f_j(\mathbf{x})$ for $\mathbf{x} \in D$ is a bounded set in \mathbb{R} .

Here is a proof of the extreme value theorem.

Theorem 15.12.2 Let C be closed and bounded and let $f : C \rightarrow \mathbb{R}$ be continuous. Then f achieves its maximum and its minimum on C . This means there exist $\mathbf{x}_1, \mathbf{x}_2 \in C$ such that for all $\mathbf{x} \in C$,

$$f(\mathbf{x}_1) \leq f(\mathbf{x}) \leq f(\mathbf{x}_2).$$

Proof: Let $M = \sup \{f(\mathbf{x}) : \mathbf{x} \in C\}$. Then there exists \mathbf{x}_n such that $f(\mathbf{x}_n) \uparrow M$. Then by compactness, there is a subsequence $\{\mathbf{x}_{n_k}\}$ such that $\mathbf{x}_{n_k} \rightarrow \mathbf{x} \in C$. It follows from continuity that $f(\mathbf{x}) = \lim_{k \rightarrow \infty} f(\mathbf{x}_{n_k}) = M$. The case for the minimum value is completely similar. Note that this shows that f is bounded. ■

As in the case of a function of one variable, there is a concept of uniform continuity.

Definition 15.12.3 A function $\mathbf{f} : D(\mathbf{f}) \rightarrow \mathbb{R}^q$ is uniformly continuous if for every $\varepsilon > 0$ there exists $\delta > 0$ such that whenever \mathbf{x}, \mathbf{y} are points of $D(\mathbf{f})$ such that $|\mathbf{x} - \mathbf{y}| < \delta$, it follows $|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| < \varepsilon$.

Theorem 15.12.4 Let $\mathbf{f} : K \rightarrow \mathbb{R}^q$ be continuous at every point of K where K is a closed and bounded (sequentially compact) set in \mathbb{R}^p . Then \mathbf{f} is uniformly continuous.

Proof: Suppose not. Then there exists $\varepsilon > 0$ and sequences $\{\mathbf{x}_j\}$ and $\{\mathbf{y}_j\}$ of points in K such that $|\mathbf{x}_j - \mathbf{y}_j| < \frac{1}{j}$ but $|\mathbf{f}(\mathbf{x}_j) - \mathbf{f}(\mathbf{y}_j)| \geq \varepsilon$. Then by Theorem 15.4.2 on Page 326 which says K is sequentially compact, there is a subsequence $\{\mathbf{x}_{n_k}\}$ of $\{\mathbf{x}_j\}$ which converges to a point $\mathbf{x} \in K$. Then since $|\mathbf{x}_{n_k} - \mathbf{y}_{n_k}| < \frac{1}{k}$, it follows that $\{\mathbf{y}_{n_k}\}$ also converges to \mathbf{x} . Therefore,

$$\varepsilon \leq \lim_{k \rightarrow \infty} |\mathbf{f}(\mathbf{x}_{n_k}) - \mathbf{f}(\mathbf{y}_{n_k})| = |\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x})| = 0,$$

a contradiction. Therefore, \mathbf{f} is uniformly continuous as claimed. ■

15.13 Convergence of Functions

There are two kinds of convergence for a sequence of functions described in the next definition, pointwise convergence and uniform convergence. Of the two, uniform convergence is far better and tends to be the kind of thing encountered in complex analysis. Pointwise convergence is more often encountered in real analysis and necessitates much more difficult theorems. Note that in so far as distance and open and closed and compact sets are concerned $\mathbb{R}^2 = \mathbb{C}$. Thus there would be no loss of generality in the following definition if \mathbb{C} were replaced with \mathbb{R} .

Definition 15.13.1 Let $S \subseteq \mathbb{C}^p$ and let $f_n : S \rightarrow \mathbb{C}^q$ for $n = 1, 2, \dots$. Then $\{f_n\}$ is said to converge pointwise to f on S if for all $x \in S$,

$$f_n(x) \rightarrow f(x)$$

for each x . The sequence is said to converge uniformly to f on S if

$$\lim_{n \rightarrow \infty} \left(\sup_{x \in S} |f_n(x) - f(x)| \right) = 0$$

$\sup_{x \in S} |f_n(x) - f(x)|$ is denoted as $\|f_n - f\|_\infty$ or just $\|f_n - f\|$ for short. $\|\cdot\|$ is called the uniform norm.

To illustrate the difference in the two types of convergence, here is a standard example shown earlier.

Example 15.13.2 Let

$$f(x) \equiv \begin{cases} 0 & \text{if } x \in [0, 1) \\ 1 & \text{if } x = 1 \end{cases}$$

Also let $f_n(x) \equiv x^n$ for $x \in [0, 1]$. Then f_n converges pointwise to f on $[0, 1]$ but does not converge uniformly to f on $[0, 1]$.

Note how the target function is not continuous although each function in the sequence is. The next theorem shows that this kind of loss of continuity **never** occurs when you have uniform convergence. The theorem holds generally when $S \subseteq X$ a normed linear space and f, f_n have values in Y another normed linear space. You should fill in the details to be sure you understand this. You simply replace $|\cdot|$ with $\|\cdot\|$ for an appropriate norm.

Theorem 15.13.3 Let $f_n : S \rightarrow \mathbb{C}^q$ be continuous and let f_n converge uniformly to f on S . Then if f_n is continuous at $x \in S$, it follows that f is also continuous at x .

Proof: Let $\varepsilon > 0$ be given. Let N be such that if $n \geq N$, then

$$\sup_{y \in S} |f_n(y) - f(y)| \equiv \|f_n - f\|_\infty < \frac{\varepsilon}{3}$$

Pick such an n . Then by continuity of f_n at x , there exists $\delta > 0$ such that if $|y - x| < \delta$, then $|f_n(y) - f_n(x)| < \frac{\varepsilon}{3}$. Then if $|y - x| < \delta$, $y \in S$, then

$$\begin{aligned} |f(x) - f(y)| &\leq |f(x) - f_n(x)| + |f_n(x) - f_n(y)| + |f_n(y) - f(y)| \\ &< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon \end{aligned}$$

Thus f is continuous at x as claimed. ■

15.14 Fundamental Theorem of Algebra

Recall from the chapter on prerequisite material the basic properties of complex numbers, complex absolute value and so forth. See Definition 1.13.3 and the following material after this definition. If you have a sequence of complex numbers $\{z_k\}$ where $z_k = x_k + iy_k$ then to say that $|z_k|$ is bounded is to say that $\sqrt{x_k^2 + y_k^2}$ is bounded. In other words, the ordered pairs (x_k, y_k) are in a bounded subset of \mathbb{R}^2 . Also to say that $\lim_{k \rightarrow \infty} |z_k - z| = 0$ is the definition of what you mean by $\lim_{k \rightarrow \infty} z_k = z$ and it is the same as saying that $\lim_{k \rightarrow \infty} (x_k, y_k) = (x, y)$ where $z = x + iy$. Thus, if you have a bounded sequence of complex numbers $\{z_k\}$, you must have $\{x_k\}$ and $\{y_k\}$ both be a bounded sequence in \mathbb{R} and so $\{x_k\}$ is contained in some interval $[a, b]$ and $\{y_k\}$ is contained in some interval $[c, d]$. Thus such a bounded sequence must have a subsequence, still denoted as z_k such that $x_k \rightarrow x \in [a, b]$ and $y_k \rightarrow y \in [c, d]$. This yields the following simple observation sometimes called the Weierstrass Bolzano theorem.

Theorem 15.14.1 *Let $\{z_k\}$ be a sequence of complex numbers such that $|z_k|$ is a bounded sequence of real numbers. Then there exists a subsequence $\{z_{n_k}\}$ and a complex number z such that $\lim_{k \rightarrow \infty} z_{n_k} = z$. Sets of the form $K \equiv \{z \in \mathbb{C} : |z| \leq r\}$ are sequentially compact.*

Proof: It only remains to verify the last assertion. Letting $\{z_k\} \subseteq K$, the above discussion shows that there exists z and a subsequence $\{z_{n_k}\}$ such that $z_{n_k} \rightarrow z$. It only remains to verify that $z \in K$. However, this is clear from the triangle inequality. Indeed,

$$|z| \leq |z - z_k| + |z_k| \leq |z - z_k| + r$$

Hence,

$$|z| \leq \lim_{k \rightarrow \infty} |z - z_{n_k}| + r = r. \blacksquare$$

Lemma 15.14.2 *Every polynomial $p(z)$ having complex coefficients is continuous. That is, if $z_k \rightarrow z$, then $p(z_k) \rightarrow p(z)$.*

Proof: Recall that if $z_k = x_k + iy_k$, $z = x + iy$, the convergence of z_k to z is equivalent to convergence of x_k to x and convergence of y_k to y . Also,

$$z^n = (x + iy)^n = \sum_{j=1}^n \binom{n}{j} (i)^j x^{n-j} y^j$$

Thus breaking into real and imaginary parts,

$$p(z) = \operatorname{Re} p(z) + i \operatorname{Im} p(z)$$

and each of $\operatorname{Re} p(z)$ and $\operatorname{Im} p(z)$ are polynomials in x and y as defined in Definition 15.7.2. Therefore, these are each continuous functions of (x, y) by Theorem 15.7.1 and as

$$x_k \rightarrow x, y_k \rightarrow y$$

it follows that $\operatorname{Re} p(z_k) \rightarrow \operatorname{Re} p(z)$, $\operatorname{Im} p(z_k) \rightarrow \operatorname{Im} p(z)$. \blacksquare

Theorem 15.14.3 *Let $p(z)$ be a polynomial of degree $n \geq 1$ having complex coefficients. Then there exists z_0 such that $p(z_0) = 0$, a zero of the polynomial.*

Proof: Suppose the nonconstant polynomial

$$p(z) = a_0 + a_1z + \cdots + a_nz^n, a_n \neq 0,$$

has no zero in \mathbb{C} . By the triangle inequality,

$$\begin{aligned} |p(z)| &\geq |a_n||z|^n - |a_0 + a_1z + \cdots + a_{n-1}z^{n-1}| \\ &\geq |a_n||z|^n - (|a_0| + |a_1||z| + \cdots + |a_{n-1}||z|^{n-1}) \end{aligned}$$

Now the term $|a_n||z|^n$ dominates all the other terms which have $|z|$ raised to a lower power and so $\lim_{|z| \rightarrow \infty} |p(z)| = \infty$. Now let

$$0 \leq \lambda \equiv \inf\{|p(z)| : z \in \mathbb{C}\}$$

Then since $\lim_{|z| \rightarrow \infty} |p(z)| = \infty$, it follows that there exists $r > 0$ such that if $|z| > r$, then $|p(z)| \geq 1 + \lambda$. It follows that

$$\lambda = \inf\{|p(z)| : |z| \leq r\}$$

Since $K \equiv \{z : |z| \leq r\}$ is sequentially compact, it follows that, letting $\{z_k\} \subseteq K$ with $|p(z_k)| \leq \lambda + 1/k$, there is a subsequence still denoted as $\{z_k\}$ such that $\lim_{k \rightarrow \infty} z_k = z_0 \in K$. Then $|p(z_0)| = \lambda$ and so $\lambda > 0$. Thus,

$$|p(z_0)| = \min_{z \in K} |p(z)| = \min_{z \in \mathbb{C}} |p(z)| > 0$$

Then let $q(z) = \frac{p(z+z_0)}{p(z_0)}$. This is also a polynomial which has no zeros and the minimum of $|q(z)|$ is 1 and occurs at $z = 0$. Since $q(0) = 1$, it follows $q(z) = 1 + a_kz^k + r(z)$ where $r(z)$ consists of higher order terms. Here a_k is the first coefficient of $q(z)$ which is nonzero. Choose a sequence, $z_n \rightarrow 0$, such that $a_kz_n^k < 0$. For example, let $-a_kz_n^k = (1/n)$. Then

$$|q(z_n)| = |1 + a_kz_n^k + r(z_n)| \leq 1 - 1/n + |r(z_n)| = 1 + a_kz_n^k + |r(z_n)| < 1$$

for all n large enough because $|r(z_n)|$ is small compared with $|a_kz_n^k|$ since it involves higher order terms. This is a contradiction. Thus there must be a zero for the original polynomial $p(z)$. ■

15.15 Exercises

1. Let $\mathbf{f}(t) = (t, t^2 + 1, \frac{t}{t+1})$ and let $\mathbf{g}(t) = (t + 1, 1, \frac{t}{t^2+1})$. Find $\mathbf{f} \cdot \mathbf{g}$.
2. Let \mathbf{f}, \mathbf{g} be given in the previous problem. Find $\mathbf{f} \times \mathbf{g}$.
3. Let $\mathbf{f}(t) = (t, t^2, t^3)$, $\mathbf{g}(t) = (1, t^2, t^2)$, and $\mathbf{h}(t) = (\sin t, t, 1)$. Find the time rate of change of the box product of the vectors \mathbf{f}, \mathbf{g} , and \mathbf{h} .
4. Let $\mathbf{f}(t) = (t, \sin t)$. Show \mathbf{f} is continuous at every point t .
5. Suppose $|\mathbf{f}(x) - \mathbf{f}(y)| \leq K|x - y|$ where K is a constant. Show that \mathbf{f} is everywhere continuous. Functions satisfying such an inequality are called Lipschitz functions.

6. Suppose $|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| \leq K|\mathbf{x} - \mathbf{y}|^\alpha$ where K is a constant and $\alpha \in (0, 1)$. Show that \mathbf{f} is everywhere continuous. Functions like this are called Hölder continuous.
7. Suppose $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ is given by $f(\mathbf{x}) = 3x_1x_2 + 2x_3^2$. Use Theorem 15.7.1 to verify that f is continuous. **Hint:** You should first verify that the function $\pi_k : \mathbb{R}^3 \rightarrow \mathbb{R}$ given by $\pi_k(\mathbf{x}) = x_k$ is a continuous function.
8. Show that if $f : \mathbb{R}^q \rightarrow \mathbb{R}$ is a polynomial then it is continuous.
9. State and prove a theorem about continuity of quotients of continuous functions.
10. Let

$$f(x, y) \equiv \begin{cases} \frac{x^2 - y^2}{x^2 + y^2} & \text{if } (x, y) \neq (0, 0) \\ 0 & \text{if } (x, y) = (0, 0) \end{cases}.$$

Find $\lim_{(x,y) \rightarrow (0,0)} f(x, y)$ if it exists. If it does not exist, tell why it does not exist.

Hint: Consider along the line $y = x$ and along the line $y = 0$.

11. Find the following limits if possible

(a) $\lim_{(x,y) \rightarrow (0,0)} \frac{x^2 - y^2}{x^2 + y^2}.$

(b) $\lim_{(x,y) \rightarrow (0,0)} \frac{x(x^2 - y^2)}{(x^2 + y^2)}.$

(c) $\lim_{(x,y) \rightarrow (0,0)} \frac{(x^2 - y^4)^2}{(x^2 + y^4)^2}.$ **Hint:** Consider along $y = 0$ and along $x = y^2$.

(d) $\lim_{(x,y) \rightarrow (0,0)} x \sin\left(\frac{1}{x^2 + y^2}\right).$

(e) $\lim_{(x,y) \rightarrow (1,2)} \frac{-2yx^2 + 8yx + 34y + 3y^3 - 18y^2 + 6x^2 - 13x - 20 - xy^2 - x^3}{-y^2 + 4y - 5 - x^2 + 2x}.$ **Hint:** Write in the variables

$$(s, t) = (x - 1, y - 2).$$

12. Suppose $\lim_{x \rightarrow 0} f(x, 0) = 0 = \lim_{y \rightarrow 0} f(0, y)$. Does it follow that

$$\lim_{(x,y) \rightarrow (0,0)} f(x, y) = 0?$$

Prove or give counter example.

13. $\mathbf{f} : D \subseteq \mathbb{R}^p \rightarrow \mathbb{R}^q$ is Lipschitz continuous or just Lipschitz for short if there exists a constant K such that

$$|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| \leq K|\mathbf{x} - \mathbf{y}|$$

for all $\mathbf{x}, \mathbf{y} \in D$. Show every Lipschitz function is uniformly continuous which means that given $\varepsilon > 0$ there exists $\delta > 0$ independent of \mathbf{x} such that if $|\mathbf{x} - \mathbf{y}| < \delta$, then $|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| < \varepsilon$.

14. If \mathbf{f} is uniformly continuous, does it follow that $|\mathbf{f}|$ is also uniformly continuous? If $|\mathbf{f}|$ is uniformly continuous does it follow that \mathbf{f} is uniformly continuous? Answer the same questions with “uniformly continuous” replaced with “continuous”. Explain why.

15. Let f be defined on the positive integers. Thus $D(f) = \mathbb{N}$. Show that f is automatically continuous at every point of $D(f)$. Is it also uniformly continuous? What does this mean about the concept of continuous functions being those which can be graphed without taking the pencil off the paper?

16. Let

$$f(x, y) = \frac{(x^2 - y^4)^2}{(x^2 + y^4)^2} \text{ if } (x, y) \neq (0, 0)$$

Show $\lim_{t \rightarrow 0} f(tx, ty) = 1$ for any choice of (x, y) . Using Problem 11c, what does this tell you about limits existing just because the limit along any line exists.

17. Let $f(x, y, z) = x^2y + \sin(xyz)$. Does f achieve a maximum on the set

$$\{(x, y, z) : x^2 + y^2 + 2z^2 \leq 8\}?$$

Explain why.

18. Suppose x is defined to be a limit point of a set A if and only if for all $r > 0$, $B(x, r)$ contains a point of A different than x . Show this is equivalent to the above definition of limit point.
19. Give an example of an infinite set of points in \mathbb{R}^3 which has no limit points. Show that if $D(f)$ equals this set, then f is continuous. Show that more generally, if f is any function for which $D(f)$ has no limit points, then f is continuous.
20. Let $\{x_k\}_{k=1}^n$ be any finite set of points in \mathbb{R}^p . Show this set has no limit points.
21. Suppose S is any set of points such that every pair of points is at least as far apart as 1. Show S has no limit points.
22. Find $\lim_{x \rightarrow 0} \frac{\sin(|x|)}{|x|}$ and prove your answer from the definition of limit.
23. Suppose g is a continuous vector valued function of one variable defined on $[0, \infty)$. Prove

$$\lim_{x \rightarrow x_0} g(|x|) = g(|x_0|).$$

24. Let $U = \{(x, y, z) \text{ such that } z > 0\}$. Determine whether U is open, closed or neither.
25. Let $U = \{(x, y, z) \text{ such that } z \geq 0\}$. Determine whether U is open, closed or neither.
26. Let $U = \{(x, y, z) \text{ such that } \sqrt{x^2 + y^2 + z^2} < 1\}$. Tell whether U is open, closed or neither.
27. Let $U = \{(x, y, z) \text{ such that } \sqrt{x^2 + y^2 + z^2} \leq 1\}$. Tell whether U is open, closed or neither.
28. Show carefully that \mathbb{R}^p is both open and closed.
29. Show that every non empty open set in \mathbb{R}^p is the union of open balls contained in it.
30. Show the intersection of any two open sets is an open set.

31. Closed sets were defined to be those sets which are complements of open sets. Show that a set is closed if and only if it contains all its limit points.
32. Prove the extreme value theorem, a continuous function achieves its maximum and minimum on any closed and bounded set C . **Hint:** Suppose $\lambda = \sup \{f(x) : x \in C\}$. Then there exists $\{x_n\} \subseteq C$ such that $f(x_n) \rightarrow \lambda$. Now select a convergent subsequence. Do the same for the minimum.
33. If \mathcal{C} is a collection of open sets such that $\cup \mathcal{C} \supseteq H$ a closed and bounded set. A **Lebesgue number** δ is one which has the property that if $x \in H$, then $B(x, \delta)$ is contained in some set of \mathcal{C} . Show that there exists a Lebesgue number. **Hint:** If there is no Lebesgue number, then for each $n \in \mathbb{N}$, $1/n$ is not a Lebesgue number. Hence there exists $x_n \in H$ such that $B(x_n, 1/n)$ is not contained in a single set of \mathcal{C} . Extract a convergent subsequence, still denoted as $x_n \rightarrow x$. Then $B(x, \delta)$ is contained in a single set of \mathcal{C} . Isn't it the case that $B(x_n, 1/n)$ is contained in $B(x, \delta)$ for all n large enough? Isn't this a contradiction?
34. Let C be a closed and bounded set and suppose $f : C \rightarrow \mathbb{R}^m$ is continuous. Show that f must also be **uniformly continuous**. This means: For every $\varepsilon > 0$ there exists $\delta > 0$ such that whenever $x, y \in C$ and $|x - y| < \delta$, it follows $|f(x) - f(y)| < \varepsilon$. It is in the chapter but go over it again. This is a good time to review the definition of continuity so you will see the difference. **Hint:** Suppose it is not so. Then there exists $\varepsilon > 0$ and $\{x_k\}$ and $\{y_k\}$ such that $|x_k - y_k| < \frac{1}{k}$ but $|f(x_k) - f(y_k)| \geq \varepsilon$.
35. A set K is **compact** means that if \mathcal{C} is a set of open sets such that $\cup \mathcal{C} \supseteq K$, then there exists a finite subset $\{U_1, \dots, U_n\} \subseteq \mathcal{C}$ such that $\cup_{i=1}^n U_i \supseteq K$. Show every closed and bounded set K in \mathbb{R}^p is compact. (Open covers admit finite sub covers.) Next show that if a set in \mathbb{R}^p is compact, then it must be closed and bounded. This is called the Heine Borel theorem. **Hint:** To show closed and bounded is compact, you might use the technique of chopping into small pieces of the above Problem 33. You could also do something like the following. Let δ be a Lebesgue number for the open cover \mathcal{C} of K . Now consider $B(x_1, \delta)$. If it covers K you are done. Otherwise, pick x_2 not in it. Consider $B(x_2, \delta)$. If these two balls cover K , then you are done. Otherwise pick x_3 not covered. Continue this way. Argue the sequential compactness of K requires this process to stop in finitely many steps. If a set K is compact, then it obviously must be bounded. Otherwise, you could consider the open cover $\{B(x, n)\}_{n=1}^\infty$. If the set K is not closed, then there is a point not in K called x and a sequence of points $\{x_k\}$ of K converging to x . Explain why $F_m \equiv \cup_{k=m}^\infty x_k$ is closed. Consider the increasing sequence of open sets F_m^C .
36. Suppose S is a nonempty set in \mathbb{R}^p . Define

$$\text{dist}(x, S) \equiv \inf \{|x - y| : y \in S\}.$$

Show that

$$|\text{dist}(x, S) - \text{dist}(y, S)| \leq |x - y|.$$

Hint: Suppose $\text{dist}(x, S) < \text{dist}(y, S)$. If these are equal there is nothing to show. Explain why there exists $z \in S$ such that $|x - z| < \text{dist}(x, S) + \varepsilon$. Now explain why

$$|\text{dist}(x, S) - \text{dist}(y, S)| = \text{dist}(y, S) - \text{dist}(x, S) \leq |y - z| - (|x - z| - \varepsilon)$$

Now use the triangle inequality and observe that ε is arbitrary.

37. Suppose H is a closed set and $H \subseteq U \subseteq \mathbb{R}^p$, an open set. Show there exists a continuous function defined on \mathbb{R}^p , f such that $f(\mathbb{R}^p) \subseteq [0, 1]$, $f(x) = 0$ if $x \notin U$ and $f(x) = 1$ if $x \in H$. **Hint:** Try something like

$$\frac{\text{dist}(x, U^C)}{\text{dist}(x, U^C) + \text{dist}(x, H)},$$

where $U^C \equiv \mathbb{R}^p \setminus U$, a closed set. You need to explain why the denominator is never equal to zero. The rest is supplied by Problem 36. This is a special case of a major theorem called Urysohn's lemma.

Chapter 16

Space Curves

A vector valued function of one variable t traces out a curve in space. Given values of t result in various points. The resulting set of points is called a space curve. The function used to describe this set of points is called a parametrization. The curve itself is called a parametric curve.

16.1 Using MATLAB to Graph Space Curves

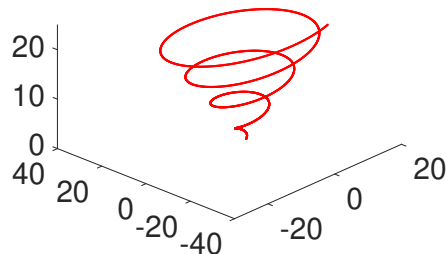
It is useful and fun to graph parametric curves if you use something like MATLAB to do the hard work. I will show you the syntax with an example.

Example 16.1.1 *Graph the space curve which has parametrization*

$$x = t \cos(t), y = t \sin(t), z = t, t \in [0, 24]$$

Here is the way you do it in MATLAB.

```
>> t=[0:.01:24];  
x=t.*cos(t);  
y=t.*sin(t);  
z=t^2;  
plot3(x,y,z,'LineWidth',2,'color','red')
```



16.2 The Derivative and Integral

The following definition is on the derivative and integral of a vector valued function of one variable.

Definition 16.2.1 *The derivative of a function $\mathbf{f}'(t)$, is defined as the following limit whenever the limit exists. If the limit does not exist, then neither does $\mathbf{f}'(t)$.*

$$\lim_{h \rightarrow 0} \frac{\mathbf{f}(t+h) - \mathbf{f}(t)}{h} \equiv \mathbf{f}'(t)$$

As before,

$$\mathbf{f}'(t) = \lim_{s \rightarrow t} \frac{\mathbf{f}(s) - \mathbf{f}(t)}{s - t}.$$

The function of h on the left is called the difference quotient just as it was for a scalar valued function. If $\mathbf{f}(t) = (f_1(t), \dots, f_p(t))$ and $\int_a^b f_i(t) dt$ exists for each $i = 1, \dots, p$, then $\int_a^b \mathbf{f}(t) dt$ is defined as the vector

$$\left(\int_a^b f_1(t) dt, \dots, \int_a^b f_p(t) dt \right).$$

This is what is meant by saying \mathbf{f} is Riemann integrable.

Here is a simple proposition which is useful to have.

Proposition 16.2.2 *Let $a \leq b$, $\mathbf{f} = (f_1, \dots, f_n)$ is vector valued and each f_i is continuous, then*

$$\left| \int_a^b \mathbf{f}(t) dt \right| \leq \sqrt{n} \int_a^b |\mathbf{f}(t)| dt.$$

Proof: This follows from the following computation.

$$\begin{aligned} \left| \int_a^b \mathbf{f}(t) dt \right| &\equiv \left| \left(\int_a^b f_1(t) dt, \dots, \int_a^b f_n(t) dt \right) \right| \\ &= \left(\sum_{i=1}^n \left| \int_a^b f_i(t) dt \right|^2 \right)^{1/2} \leq \left(\sum_{i=1}^n \left(\int_a^b |f_i(t)| dt \right)^2 \right)^{1/2} \\ &\leq \left(n \max_i \left(\int_a^b |f_i(t)| dt \right)^2 \right)^{1/2} = \sqrt{n} \max_i \left(\int_a^b |f_i(t)| dt \right) \\ &\leq \sqrt{n} \int_a^b |\mathbf{f}(t)| dt \blacksquare \end{aligned}$$

As in the case of a scalar valued function, differentiability implies continuity but not the other way around.

Theorem 16.2.3 *If $\mathbf{f}'(t)$ exists, then \mathbf{f} is continuous at t .*

Proof: Suppose $\varepsilon > 0$ is given and choose $\delta_1 > 0$ such that if $|h| < \delta_1$,

$$\left| \frac{\mathbf{f}(t+h) - \mathbf{f}(t)}{h} - \mathbf{f}'(t) \right| < 1.$$

then for such h , the triangle inequality implies $|\mathbf{f}(t+h) - \mathbf{f}(t)| < |h| + |\mathbf{f}'(t)| |h|$. Now letting $\delta < \min\left(\delta_1, \frac{\varepsilon}{1+|\mathbf{f}'(t)|}\right)$ it follows if $|h| < \delta$, then $|\mathbf{f}(t+h) - \mathbf{f}(t)| < \varepsilon$. Letting $y = h + t$, this shows that if $|y - t| < \delta$, $|\mathbf{f}(y) - \mathbf{f}(t)| < \varepsilon$ which proves \mathbf{f} is continuous at t . ■

As in the scalar case, there is a fundamental theorem of calculus.

Theorem 16.2.4 If $\mathbf{f} \in R([a, b])$ and if \mathbf{f} is continuous at $t \in (a, b)$, then

$$\frac{d}{dt} \left(\int_a^t \mathbf{f}(s) ds \right) = \mathbf{f}(t).$$

Proof: Say $\mathbf{f}(t) = (f_1(t), \dots, f_p(t))$. Then it follows

$$\frac{1}{h} \int_a^{t+h} \mathbf{f}(s) ds - \frac{1}{h} \int_a^t \mathbf{f}(s) ds = \left(\frac{1}{h} \int_t^{t+h} f_1(s) ds, \dots, \frac{1}{h} \int_t^{t+h} f_p(s) ds \right)$$

and $\lim_{h \rightarrow 0} \frac{1}{h} \int_t^{t+h} f_i(s) ds = f_i(t)$ for each $i = 1, \dots, p$ from the fundamental theorem of calculus for scalar valued functions. Therefore,

$$\lim_{h \rightarrow 0} \frac{1}{h} \int_a^{t+h} \mathbf{f}(s) ds - \frac{1}{h} \int_a^t \mathbf{f}(s) ds = (f_1(t), \dots, f_p(t)) = \mathbf{f}(t). \blacksquare$$

Example 16.2.5 Let $\mathbf{f}(x) = \mathbf{c}$ where \mathbf{c} is a constant. Find $\mathbf{f}'(x)$.

The difference quotient,

$$\frac{\mathbf{f}(x+h) - \mathbf{f}(x)}{h} = \frac{\mathbf{c} - \mathbf{c}}{h} = \mathbf{0}$$

Therefore,

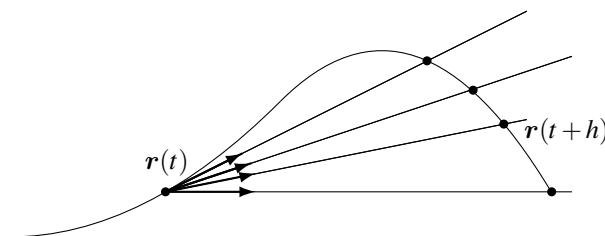
$$\lim_{h \rightarrow 0} \frac{\mathbf{f}(x+h) - \mathbf{f}(x)}{h} = \lim_{h \rightarrow 0} \mathbf{0} = \mathbf{0}$$

Example 16.2.6 Let $\mathbf{f}(t) = (at, bt)$ where a, b are constants. Find $\mathbf{f}'(t)$.

From the above discussion this derivative is just the vector valued functions whose components consist of the derivatives of the components of \mathbf{f} . Thus $\mathbf{f}'(t) = (a, b)$.

16.2.1 Geometric and Physical Significance of the Derivative

Suppose \mathbf{r} is a vector valued function of a parameter t not necessarily time and consider the following picture of the points traced out by \mathbf{r} .



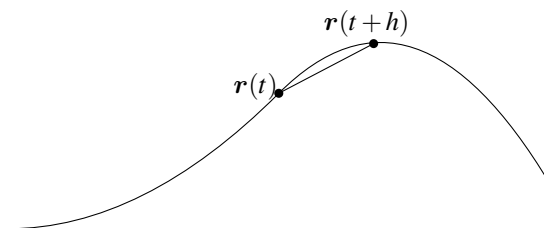
In this picture there are unit vectors in the direction of the vector from $\mathbf{r}(t)$ to $\mathbf{r}(t+h)$. You can see that it is reasonable to suppose these unit vectors, if they converge, converge to a unit vector \mathbf{T} which is tangent to the curve at the point $\mathbf{r}(t)$. Now each of these unit vectors is of the form

$$\frac{\mathbf{r}(t+h) - \mathbf{r}(t)}{|\mathbf{r}(t+h) - \mathbf{r}(t)|} \equiv \mathbf{T}_h.$$

Thus $\mathbf{T}_h \rightarrow \mathbf{T}$, a unit tangent vector to the curve at the point $\mathbf{r}(t)$. Therefore,

$$\begin{aligned} \mathbf{r}'(t) &\equiv \lim_{h \rightarrow 0} \frac{\mathbf{r}(t+h) - \mathbf{r}(t)}{h} = \lim_{h \rightarrow 0} \frac{|\mathbf{r}(t+h) - \mathbf{r}(t)|}{h} \frac{\mathbf{r}(t+h) - \mathbf{r}(t)}{|\mathbf{r}(t+h) - \mathbf{r}(t)|} \\ &= \lim_{h \rightarrow 0} \frac{|\mathbf{r}(t+h) - \mathbf{r}(t)|}{h} \mathbf{T}_h = |\mathbf{r}'(t)| \mathbf{T}. \end{aligned}$$

In the case that t is time, the expression $|\mathbf{r}(t+h) - \mathbf{r}(t)|$ is a good approximation for the distance traveled by the object on the time interval $[t, t+h]$. The real distance would be the length of the curve joining the two points but if h is very small, this is essentially equal to $|\mathbf{r}(t+h) - \mathbf{r}(t)|$ as suggested by the picture below.



Therefore, $\frac{|\mathbf{r}(t+h) - \mathbf{r}(t)|}{h}$ gives for small h , the approximate distance travelled on the time interval $[t, t+h]$ divided by the length of time h . Therefore, this expression is really the average speed of the object on this small time interval and so the limit as $h \rightarrow 0$, deserves to be called the instantaneous speed of the object. Thus $|\mathbf{r}'(t)| \mathbf{T}$ represents the speed times a unit direction vector \mathbf{T} which defines the direction in which the object is moving. Thus $\mathbf{r}'(t)$ is the velocity of the object. This is the physical significance of the derivative when t is time. In general, $\mathbf{r}'(t)$ and $\mathbf{T}(t)$ are vectors tangent to the curve which point in the direction of motion.

How do you go about computing $\mathbf{r}'(t)$? Letting $\mathbf{r}(t) = (r_1(t), \dots, r_q(t))$, the expression

$$\frac{\mathbf{r}(t_0+h) - \mathbf{r}(t_0)}{h} \tag{16.1}$$

is equal to

$$\left(\frac{r_1(t_0+h) - r_1(t_0)}{h}, \dots, \frac{r_q(t_0+h) - r_q(t_0)}{h} \right).$$

Then as h converges to 0, [16.1](#) converges to $\mathbf{v} \equiv (v_1, \dots, v_q)$ where $v_k = r'_k(t)$. This is because of Theorem [15.8.6](#) on Page [330](#), which says that the term in [16.1](#) gets close to a vector \mathbf{v} if and only if all the coordinate functions of the term in [16.1](#) get close to the corresponding coordinate functions of \mathbf{v} .

In the case where t is time, this simply says the velocity vector equals the vector whose components are the derivatives of the components of the displacement vector $\mathbf{r}(t)$.

Example 16.2.7 Let $\mathbf{r}(t) = (\sin t, t^2, t+1)$ for $t \in [0, 5]$. Find a tangent line to the curve parameterized by \mathbf{r} at the point $\mathbf{r}(2)$.

From the above discussion, a direction vector has the same direction as $\mathbf{r}'(2)$. Therefore, it suffices to simply use $\mathbf{r}'(2)$ as a direction vector for the line. $\mathbf{r}'(2) = (\cos 2, 4, 1)$. Therefore, a parametric equation for the tangent line is

$$(\sin 2, 4, 3) + t(\cos 2, 4, 1) = (x, y, z).$$

Example 16.2.8 Let $\mathbf{r}(t) = (\sin t, t^2, t+1)$ for $t \in [0, 5]$. Find the velocity vector when $t = 1$.

From the above discussion, this is simply $\mathbf{r}'(1) = (\cos 1, 2, 1)$.

16.2.2 Differentiation Rules

There are rules which relate the derivative to the various operations done with vectors such as the dot product, the cross product, vector addition, and scalar multiplication.

Theorem 16.2.9 Let $a, b \in \mathbb{R}$ and suppose $\mathbf{f}'(t)$ and $\mathbf{g}'(t)$ exist. Then the following formulas are valid.

$$(a\mathbf{f} + b\mathbf{g})'(t) = a\mathbf{f}'(t) + b\mathbf{g}'(t). \quad (16.2)$$

$$(\mathbf{f} \cdot \mathbf{g})'(t) = \mathbf{f}'(t) \cdot \mathbf{g}(t) + \mathbf{f}(t) \cdot \mathbf{g}'(t) \quad (16.3)$$

If \mathbf{f}, \mathbf{g} have values in \mathbb{R}^3 , then

$$(\mathbf{f} \times \mathbf{g})'(t) = \mathbf{f}(t) \times \mathbf{g}'(t) + \mathbf{f}'(t) \times \mathbf{g}(t) \quad (16.4)$$

The formulas, [16.3](#), and [16.4](#) are referred to as the product rule.

Proof: The first formula is left for you to prove. Consider the second, [16.3](#).

$$\begin{aligned} & \lim_{h \rightarrow 0} \frac{\mathbf{f} \cdot \mathbf{g}(t+h) - \mathbf{f} \cdot \mathbf{g}(t)}{h} \\ &= \lim_{h \rightarrow 0} \frac{\mathbf{f}(t+h) \cdot \mathbf{g}(t+h) - \mathbf{f}(t+h) \cdot \mathbf{g}(t)}{h} + \frac{\mathbf{f}(t+h) \cdot \mathbf{g}(t) - \mathbf{f}(t) \cdot \mathbf{g}(t)}{h} \\ &= \lim_{h \rightarrow 0} \left(\mathbf{f}(t+h) \cdot \frac{(\mathbf{g}(t+h) - \mathbf{g}(t))}{h} + \frac{(\mathbf{f}(t+h) - \mathbf{f}(t))}{h} \cdot \mathbf{g}(t) \right) \\ &= \lim_{h \rightarrow 0} \sum_{k=1}^n f_k(t+h) \frac{(g_k(t+h) - g_k(t))}{h} + \sum_{k=1}^n \frac{(f_k(t+h) - f_k(t))}{h} g_k(t) \\ &= \sum_{k=1}^n f_k(t) g'_k(t) + \sum_{k=1}^n f'_k(t) g_k(t) = \mathbf{f}'(t) \cdot \mathbf{g}(t) + \mathbf{f}(t) \cdot \mathbf{g}'(t). \end{aligned}$$

Formula 16.4 is left as an exercise which follows from the product rule and the definition of the cross product. ■

Example 16.2.10 Let $\mathbf{r}(t) = (t^2, \sin t, \cos t)$ and let

$$\mathbf{p}(t) = (t, \ln(t+1), 2t).$$

Find $(\mathbf{r}(t) \times \mathbf{p}(t))'$.

From 16.4 this equals

$$(2t, \cos t, -\sin t) \times (t, \ln(t+1), 2t) + (t^2, \sin t, \cos t) \times \left(1, \frac{1}{t+1}, 2\right)$$

Example 16.2.11 Let $\mathbf{r}(t) = (t^2, \sin t, \cos t)$ Find $\int_0^\pi \mathbf{r}(t) dt$.

This equals $(\int_0^\pi t^2 dt, \int_0^\pi \sin t dt, \int_0^\pi \cos t dt) = (\frac{1}{3}\pi^3, 2, 0)$.

Example 16.2.12 An object has position

$$\mathbf{r}(t) = \left(t^3, \frac{t}{1+t}, \sqrt{t^2+2}\right)$$

kilometers where t is given in hours. Find the velocity of the object in kilometers per hour when $t = 1$.

Recall the velocity at time t was $\mathbf{r}'(t)$. Therefore, find $\mathbf{r}'(t)$ and plug in $t = 1$ to find the velocity.

$$\mathbf{r}'(t) = \left(3t^2, \frac{1(1+t)-t}{(1+t)^2}, \frac{1}{2}(t^2+2)^{-1/2} 2t\right) = \left(3t^2, \frac{1}{(1+t)^2}, \frac{1}{\sqrt{t^2+2}}t\right)$$

When $t = 1$, the velocity is

$$\mathbf{r}'(1) = \left(3, \frac{1}{4}, \frac{1}{\sqrt{3}}\right) \text{ kilometers per hour.}$$

Obviously, this can be continued. That is, you can consider the possibility of taking the derivative of the derivative and then the derivative of that and so forth. The main thing to consider about this is the notation, and it is exactly like it was in the case of a scalar valued function presented earlier. Thus $\mathbf{r}''(t)$ denotes the second derivative.

When you are given a vector valued function of one variable, sometimes it is possible to give a simple description of the curve which results. Usually it is not possible to do this!

Example 16.2.13 Describe the curve which results from the vector valued function $\mathbf{r}(t) = (\cos 2t, \sin 2t, t)$ where $t \in \mathbb{R}$.

The first two components indicate that for $\mathbf{r}(t) = (x(t), y(t), z(t))$, the pair, $(x(t), y(t))$ traces out a circle. While it is doing so, $z(t)$ is moving at a steady rate in the positive direction. Therefore, the curve which results is a cork screw shaped thing called a helix.

As an application of the theorems for differentiating curves, here is an interesting application. It is also a situation where the curve can be identified as something familiar.

Example 16.2.14 *Sound waves have the angle of incidence equal to the angle of reflection. Suppose you are in a large room and you make a sound. The sound waves spread out and you would expect your sound to be inaudible very far away. But what if the room were shaped so that the sound is reflected off the wall toward a single point, possibly far away from you? Then you might have the interesting phenomenon of someone far away hearing what you said quite clearly. How should the room be designed?*

Suppose you are located at the point P_0 and the point where your sound is to be reflected is P_1 . Consider a plane which contains the two points and let $\mathbf{r}(t)$ denote a parametrization of the intersection of this plane with the walls of the room. Then the condition that the angle of reflection equals the angle of incidence reduces to saying the angle between $P_0 - \mathbf{r}(t)$ and $-\mathbf{r}'(t)$ equals the angle between $P_1 - \mathbf{r}(t)$ and $\mathbf{r}'(t)$. Draw a picture to see this. Therefore,

$$\frac{(P_0 - \mathbf{r}(t)) \cdot (-\mathbf{r}'(t))}{|P_0 - \mathbf{r}(t)| |\mathbf{r}'(t)|} = \frac{(P_1 - \mathbf{r}(t)) \cdot (\mathbf{r}'(t))}{|P_1 - \mathbf{r}(t)| |\mathbf{r}'(t)|}.$$

This reduces to

$$\frac{(\mathbf{r}(t) - P_0) \cdot (-\mathbf{r}'(t))}{|\mathbf{r}(t) - P_0|} = \frac{(\mathbf{r}(t) - P_1) \cdot (\mathbf{r}'(t))}{|\mathbf{r}(t) - P_1|} \quad (16.5)$$

Now

$$\frac{(\mathbf{r}(t) - P_1) \cdot (\mathbf{r}'(t))}{|\mathbf{r}(t) - P_1|} = \frac{d}{dt} |\mathbf{r}(t) - P_1|$$

and a similar formula holds for P_1 replaced with P_0 . This is because

$$|\mathbf{r}(t) - P_1| = \sqrt{(\mathbf{r}(t) - P_1) \cdot (\mathbf{r}(t) - P_1)}$$

and so using the chain rule and product rule,

$$\begin{aligned} \frac{d}{dt} |\mathbf{r}(t) - P_1| &= \frac{1}{2} ((\mathbf{r}(t) - P_1) \cdot (\mathbf{r}(t) - P_1))^{-1/2} 2 ((\mathbf{r}(t) - P_1) \cdot \mathbf{r}'(t)) \\ &= \frac{(\mathbf{r}(t) - P_1) \cdot (\mathbf{r}'(t))}{|\mathbf{r}(t) - P_1|}. \end{aligned}$$

Therefore, from 16.5,

$$\frac{d}{dt} (|\mathbf{r}(t) - P_1|) + \frac{d}{dt} (|\mathbf{r}(t) - P_0|) = 0$$

showing that $|\mathbf{r}(t) - P_1| + |\mathbf{r}(t) - P_0| = C$ for some constant C . This implies the curve of intersection of the plane with the room is an ellipse having P_0 and P_1 as the foci.

16.2.3 Leibniz's Notation

Leibniz's notation also generalizes routinely. For example, $\frac{dy}{dt} = \mathbf{y}'(t)$ with other similar notations holding.

16.3 Arc Length and Orientations

The application of the integral considered here is the concept of the **length of a curve**.

Definition 16.3.1 C is a **smooth curve** in \mathbb{R}^n if there exists an interval $[a, b] \subseteq \mathbb{R}$ and functions $x_i : [a, b] \rightarrow \mathbb{R}$ such that the following conditions hold

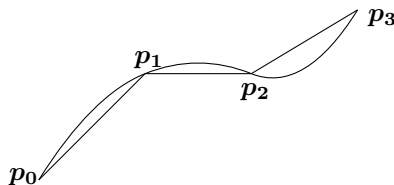
1. x_i is continuous on $[a, b]$.
2. x'_i exists and is continuous and bounded on $[a, b]$, with $x'_i(a)$ defined as the derivative from the right,

$$\lim_{h \rightarrow 0+} \frac{x_i(a+h) - x_i(a)}{h},$$

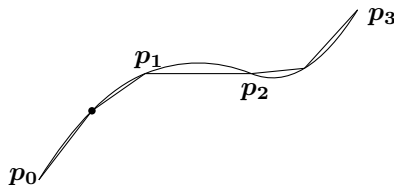
and $x'_i(b)$ defined similarly as the derivative from the left.

3. For $\mathbf{p}(t) \equiv (x_1(t), \dots, x_n(t))$, $t \rightarrow \mathbf{p}(t)$ is one to one on (a, b) .
4. $|\mathbf{p}'(t)| \equiv \left(\sum_{i=1}^n |x'_i(t)|^2 \right)^{1/2} \neq 0$ for all $t \in [a, b]$.
5. $C = \cup \{ (x_1(t), \dots, x_n(t)) : t \in [a, b] \}$.

The functions $x_i(t)$, defined above are giving the coordinates of a point in \mathbb{R}^n and the list of these functions is called a **parametrization** for the smooth curve. Note the natural direction of the interval also gives a direction for moving along the curve. Such a direction is called an **orientation**. The integral is used to define what is meant by the length of such a smooth curve. Consider such a smooth curve having parametrization (x_1, \dots, x_n) . Forming a partition of $[a, b]$, $a = t_0 < \dots < t_m = b$ and letting $\mathbf{p}_i = (x_1(t_i), \dots, x_n(t_i))$, you could consider the polygon formed by lines from \mathbf{p}_0 to \mathbf{p}_1 and from \mathbf{p}_1 to \mathbf{p}_2 and from \mathbf{p}_2 to \mathbf{p}_3 etc. to be an approximation to the curve C . The following picture illustrates what is meant by this.



Now consider what happens when the partition is refined by including more points. You can see from the following picture that the polygonal approximation would appear to be even better and that as more points are added in the partition, the sum of the lengths of the line segments seems to get close to something which deserves to be defined as the length of the curve C .



Thus the length of the curve is approximately equal to

$$\sum_{k=1}^m |\mathbf{p}(t_k) - \mathbf{p}(t_{k-1})|$$

Since the functions in the parametrization are differentiable, this is approximately

$$\sum_{k=1}^m |\mathbf{p}'(t_{k-1})| (t_k - t_{k-1})$$

which is seen to be a Riemann sum for the integral $\int_a^b |\mathbf{p}'(t)| dt$ and it is this integral which is **defined** as the length of the curve.

Definition 16.3.2 Let $\mathbf{p}(t)$, $t \in [a, b]$ be a parametrization for a smooth curve. Then the length of this curve is defined as $\int_a^b |\mathbf{p}'(t)| dt$.

Would the same length be obtained if another parametrization were used? This is a very important question because the length of the curve should depend only on the curve itself and not on the method used to trace out the curve. The answer to this question is that the length of the curve does not depend on parametrization. The proof is somewhat technical so is given later.

Does the definition of length given above correspond to the usual definition of length in the case when the curve is a line segment? It is easy to see that it does so by considering two points in \mathbb{R}^n \mathbf{p} and \mathbf{q} . A parametrization for the line segment joining these two points is

$$f_i(t) \equiv t p_i + (1-t) q_i, \quad t \in [0, 1].$$

Using the definition of length of a smooth curve just given, the length according to this definition is

$$\int_0^1 \left(\sum_{i=1}^n (p_i - q_i)^2 \right)^{1/2} dt = |\mathbf{p} - \mathbf{q}|.$$

Thus this new definition which is valid for smooth curves which may not be straight line segments gives the usual length for straight line segments.

The proof that curve length is well defined for a smooth curve contains a result which deserves to be stated as a corollary. It is proved in Lemma 16.4.6 on Page 355 but the proof is mathematically fairly difficult so it is presented later. See also Theorem 16.4.7 for the proof that length does not depend on parametrization.

Corollary 16.3.3 Let C be a smooth curve and let $\mathbf{f} : [a, b] \rightarrow C$ and $\mathbf{g} : [c, d] \rightarrow C$ be two parameterizations satisfying 1 - 5. Then $\mathbf{g}^{-1} \circ \mathbf{f}$ is either strictly increasing or strictly decreasing.

Definition 16.3.4 If $\mathbf{g}^{-1} \circ \mathbf{f}$ is increasing, then \mathbf{f} and \mathbf{g} are said to be equivalent parameterizations and this is written as $\mathbf{f} \sim \mathbf{g}$. It is also said that the two parameterizations give the same orientation for the curve when $\mathbf{f} \sim \mathbf{g}$. The symbol \sim is for the word “similar”.

When the parameterizations are equivalent, they preserve the direction of motion along the curve, and this also shows there are exactly two orientations of the curve since either $g^{-1} \circ f$ is increasing or it is decreasing. This is not hard to believe. In simple language, the message is that there are exactly two directions of motion along a curve. The difficulty is in proving this is actually the case.

Lemma 16.3.5 *The following hold for \sim .*

$$f \sim f; \quad (16.6)$$

$$\text{If } f \sim g \text{ then } g \sim f; \quad (16.7)$$

$$\text{If } f \sim g \text{ and } g \sim h, \text{ then } f \sim h. \quad (16.8)$$

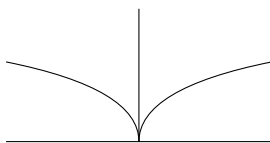
Proof: Formula 16.6 is obvious because $f^{-1} \circ f(t) = t$ so it is clearly an increasing function. If $f \sim g$ then $f^{-1} \circ g$ is increasing. Now $g^{-1} \circ f$ must also be increasing because it is the inverse of $f^{-1} \circ g$. This verifies 16.7. To see 16.8, $f^{-1} \circ h = (f^{-1} \circ g) \circ (g^{-1} \circ h)$ and so since both of these functions are increasing, it follows $f^{-1} \circ h$ is also increasing. ■

The symbol \sim is called an equivalence relation. If C is such a smooth curve just described, and if $f : [a, b] \rightarrow C$ is a parametrization of C , consider $g(t) \equiv f((a+b)-t)$, also a parametrization of C . Now by Corollary 16.3.3, if h is a parametrization, then if $f^{-1} \circ h$ is not increasing, it must be the case that $g^{-1} \circ h$ is increasing. Consequently, either $h \sim g$ or $h \sim f$. These parameterizations, h , which satisfy $h \sim f$ are called the equivalence class determined by f and those h which are similar to g are called the equivalence class determined by g . These two classes are called **orientations** of C . They give the direction of motion on C . You see that going from f to g corresponds to tracing out the curve in the opposite direction.

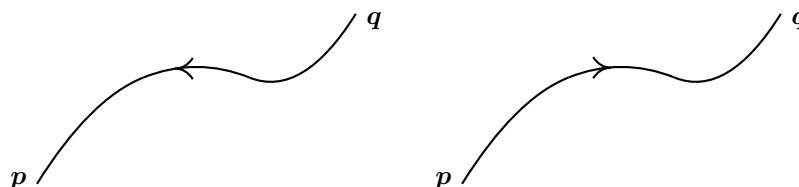
Sometimes people wonder why it is required, in the definition of a smooth curve that $p'(t) \neq 0$. Imagine t is time and $p(t)$ gives the location of a point in space. If $p'(t)$ is allowed to equal zero, the point can stop and change directions abruptly, producing a pointy place in C . Here is an example.

Example 16.3.6 *Graph the curve (t^3, t^2) for $t \in [-1, 1]$.*

In this case, $t = x^{1/3}$ and so $y = x^{2/3}$. Thus the graph of this curve looks like the picture below. Note the pointy place. Such a curve should not be considered smooth.



So what is the thing to remember from all this? First, there are certain conditions which must be satisfied for a curve to be smooth. These are listed above. Next, if you have any curve, there are two directions you can move over this curve, each called an orientation. This is illustrated in the following picture.



Either you move from p to q or you move from q to p .

Definition 16.3.7 A curve C is *piecewise smooth* if there exist points on this curve p_0, p_1, \dots, p_n such that, denoting $C_{p_{k-1}p_k}$ the part of the curve joining p_{k-1} and p_k , it follows $C_{p_{k-1}p_k}$ is a smooth curve and $\bigcup_{k=1}^n C_{p_{k-1}p_k} = C$. In other words, it is piecewise smooth if it consists of a finite number of smooth curves linked together.

Note that Example 16.3.6 is an example of a piecewise smooth curve although it is not smooth.

16.4 Arc Length and Parametrizations*



Recall that if $p(t) : t \in [a, b]$ was a parametrization of a smooth curve C , the length of C is defined as $\int_a^b |p'(t)| dt$. If some other parametrization were used to trace out C , would the same answer be obtained? To answer this question in a satisfactory manner requires some hard calculus.

16.4.1 Hard Calculus

Recall Theorem 4.0.8 about continuity and convergent sequences. It said roughly that a function f is continuous if and only if it takes convergent sequences to convergent sequences.

This next lemma was proved earlier as an application of the intermediate value theorem. I am stating it here again for convenience.

Lemma 16.4.1 Let $\phi : [a, b] \rightarrow \mathbb{R}$ be a continuous function and suppose ϕ is 1-1 on (a, b) . Then ϕ is either strictly increasing or strictly decreasing on $[a, b]$. Furthermore, ϕ^{-1} is continuous.

Corollary 16.4.2 Let $f : (a, b) \rightarrow \mathbb{R}$ be one to one and continuous. Then $f(a, b)$ is an open interval (c, d) and $f^{-1} : (c, d) \rightarrow (a, b)$ is continuous.

Proof: Since f is either strictly increasing or strictly decreasing, it follows that $f(a, b)$ is an open interval (c, d) . Assume f is decreasing. Now let $x \in (a, b)$. Why is f^{-1} is continuous at $f(x)$? Let $\varepsilon > 0$ be given. Let $\varepsilon > \eta > 0$ and $(x - \eta, x + \eta) \subseteq (a, b)$. Then $f(x) \in (f(x + \eta), f(x - \eta))$. Let

$$\delta = \min(f(x) - f(x + \eta), f(x - \eta) - f(x)).$$

Then if $|f(z) - f(x)| < \delta$, it follows

$$z \equiv f^{-1}(f(z)) \in (x - \eta, x + \eta) \subseteq (x - \varepsilon, x + \varepsilon)$$

which implies

$$|f^{-1}(f(z)) - x| = |f^{-1}(f(z)) - f^{-1}(f(x))| < \varepsilon.$$

This proves the theorem in the case where f is strictly decreasing. The case where f is increasing is similar. ■

Theorem 16.4.3 *Let $f : [a, b] \rightarrow \mathbb{R}$ be continuous and one to one. Suppose $f'(x_1)$ exists for some $x_1 \in [a, b]$ and $f'(x_1) \neq 0$. Then $(f^{-1})'(f(x_1))$ exists and is given by the formula $(f^{-1})'(f(x_1)) = \frac{1}{f'(x_1)}$.*

Proof: By Lemma 16.4.1 f is either strictly increasing or strictly decreasing and f^{-1} is continuous on $[a, b]$. Therefore there exists $\eta > 0$ such that if $0 < |f(x_1) - f(x)| < \eta$, then

$$0 < |x_1 - x| = |f^{-1}(f(x_1)) - f^{-1}(f(x))| < \delta$$

where δ is small enough that for $0 < |x_1 - x| < \delta$,

$$\left| \frac{x - x_1}{f(x) - f(x_1)} - \frac{1}{f'(x_1)} \right| < \varepsilon.$$

It follows that if $0 < |f(x_1) - f(x)| < \eta$,

$$\left| \frac{f^{-1}(f(x)) - f^{-1}(f(x_1))}{f(x) - f(x_1)} - \frac{1}{f'(x_1)} \right| = \left| \frac{x - x_1}{f(x) - f(x_1)} - \frac{1}{f'(x_1)} \right| < \varepsilon$$

Therefore, since $\varepsilon > 0$ is arbitrary,

$$\lim_{y \rightarrow f(x_1)} \frac{f^{-1}(y) - f^{-1}(f(x_1))}{y - f(x_1)} = \frac{1}{f'(x_1)}. \quad \blacksquare$$

The following obvious corollary comes from the above by not bothering with end points.

Corollary 16.4.4 *Let $f : (a, b) \rightarrow \mathbb{R}$ be continuous and one to one. Suppose $f'(x_1)$ exists for some $x_1 \in (a, b)$ and $f'(x_1) \neq 0$. Then $(f^{-1})'(f(x_1))$ exists and is given by the formula $(f^{-1})'(f(x_1)) = \frac{1}{f'(x_1)}$.*

Proof: From the definition of the derivative and continuity of f^{-1} ,

$$\lim_{f(x) \rightarrow f(x_1)} \frac{f^{-1}(f(x)) - f^{-1}(f(x_1))}{f(x) - f(x_1)} = \lim_{x \rightarrow x_1} \frac{x - x_1}{f(x) - f(x_1)} = \frac{1}{f'(x_1)}. \quad \blacksquare$$

16.4.2 Independence of Parametrization

Theorem 16.4.5 *Let $\phi : [a, b] \rightarrow [c, d]$ be one to one and suppose ϕ' exists and is continuous on $[a, b]$. Then if f is a continuous function defined on $[c, d]$*

$$\int_c^d f(s) ds = \int_a^b f(\phi(t)) |\phi'(t)| dt$$

Proof: Let $F'(s) = f(s)$. (For example, let $F(s) = \int_a^s f(r) dr$.) Then the first integral equals $F(d) - F(c)$ by the fundamental theorem of calculus. Since ϕ is one to one, it follows from Lemma 16.4.1 above that ϕ is either strictly increasing or strictly decreasing. Suppose ϕ is strictly decreasing. Then $\phi(a) = d$ and $\phi(b) = c$. Therefore, $\phi' \leq 0$ and the second integral equals

$$-\int_a^b f(\phi(t)) \phi'(t) dt = \int_b^a \frac{d}{dt} (F(\phi(t))) dt = F(\phi(a)) - F(\phi(b)) = F(d) - F(c).$$

The case when ϕ is increasing is similar but easier. ■

Lemma 16.4.6 *Let $f : [a, b] \rightarrow C$, $g : [c, d] \rightarrow C$ be parameterizations of a smooth curve which satisfy conditions 1 - 5. Then $\phi(t) \equiv g^{-1} \circ f(t)$ is 1 - 1 on (a, b) , continuous on $[a, b]$, and either strictly increasing or strictly decreasing on $[a, b]$.*

Proof: It is obvious ϕ is 1 - 1 on (a, b) from the conditions f and g satisfy. It only remains to verify continuity on $[a, b]$ because then the final claim follows from Lemma 16.4.1. If ϕ is not continuous on $[a, b]$, then there exists a sequence, $\{t_n\} \subseteq [a, b]$ such that $t_n \rightarrow t$ but $\phi(t_n)$ fails to converge to $\phi(t)$. Therefore, for some $\varepsilon > 0$, there exists a subsequence, still denoted by n such that $|\phi(t_n) - \phi(t)| \geq \varepsilon$. By sequential compactness of $[c, d]$, there is a further subsequence, still denoted by n , such that $\{\phi(t_n)\}$ converges to a point s , of $[c, d]$ which is not equal to $\phi(t)$. Thus $g^{-1} \circ f(t_n) \rightarrow s$ while $t_n \rightarrow t$. Therefore, the continuity of f and g imply $f(t_n) \rightarrow f(t)$ and $f(t_n) \rightarrow g(s)$. Thus, $g(s) = f(t)$, so $s = g^{-1} \circ f(t) = \phi(t)$, a contradiction. Therefore, ϕ is continuous as claimed. ■

Theorem 16.4.7 *The length of a smooth curve is not dependent on which parametrization is used.*

Proof: Let C be the curve and suppose $f : [a, b] \rightarrow C$ and $g : [c, d] \rightarrow C$ both satisfy conditions 1 - 5. Is it true that $\int_a^b |f'(t)| dt = \int_c^d |g'(s)| ds$?

Let $\phi(t) \equiv g^{-1} \circ f(t)$ for $t \in [a, b]$. I want to show that ϕ is C^1 on an interval of the form $[a + \delta, b - \delta]$. By the above lemma, ϕ is either strictly increasing or strictly decreasing on $[a, b]$. Suppose for the sake of simplicity that it is strictly increasing. The decreasing case is handled similarly.

Let $s_0 \in \phi([a + \delta, b - \delta]) \subset (c, d)$. Then by assumption 4 for smooth curves, $g'_i(s_0) \neq 0$ for some i . By continuity of g'_i , it follows $g'_i(s) \neq 0$ for all $s \in I$ where I is an open interval contained in $[c, d]$ which contains s_0 . It follows from the mean value theorem that on this interval g_i is either strictly increasing or strictly decreasing. Therefore, $J \equiv g_i(I)$ is also an open interval and you can define a differentiable function $h_i : J \rightarrow I$ by

$$h_i(g_i(s)) = s.$$

This implies that for $s \in I$,

$$h'_i(g_i(s)) = \frac{1}{g'_i(s)}. \quad (16.9)$$

Now letting $s = \phi(t)$ for $s \in I$, it follows $t \in J_1$, an open interval. Also, for s and t related this way, $\mathbf{f}(t) = \mathbf{g}(s)$ and so in particular, for $s \in I$, $g_i(s) = f_i(t)$. Consequently,

$$s = h_i(g_i(s)) = h_i(f_i(t)) = \phi(t)$$

and so, for $t \in J_1$,

$$\phi'(t) = h'_i(f_i(t)) f'_i(t) = h'_i(g_i(s)) f'_i(t) = \frac{f'_i(t)}{g'_i(\phi(t))} \quad (16.10)$$

which shows that ϕ' exists and is continuous on J_1 , an open interval containing $\phi^{-1}(s_0)$. Since s_0 is arbitrary, this shows ϕ' exists on $[a + \delta, b - \delta]$ and is continuous there.

Now $\mathbf{f}(t) = \mathbf{g} \circ (\mathbf{g}^{-1} \circ \mathbf{f})(t) = \mathbf{g}(\phi(t))$, and it was just shown that ϕ' is a continuous function on $[a - \delta, b + \delta]$. It follows from the chain rule applied to the components that $\mathbf{f}'(t) = \mathbf{g}'(\phi(t)) \phi'(t)$ and so, by Theorem 16.4.5,

$$\int_{\phi(a+\delta)}^{\phi(b-\delta)} |\mathbf{g}'(s)| ds = \int_{a+\delta}^{b-\delta} |\mathbf{g}'(\phi(t))| |\phi'(t)| dt = \int_{a+\delta}^{b-\delta} |\mathbf{f}'(t)| dt.$$

Now using the continuity of ϕ , \mathbf{g}' , and \mathbf{f}' on $[a, b]$ and letting $\delta \rightarrow 0+$ in the above, yields

$$\int_c^d |\mathbf{g}'(s)| ds = \int_a^b |\mathbf{f}'(t)| dt. \blacksquare$$

16.5 Exercises

1. Find the following limits if possible

(a) $\lim_{x \rightarrow 0+} \left(\frac{|x|}{x}, \sin x/x, \cos x \right)$

(b) $\lim_{x \rightarrow 0+} \left(\frac{x}{|x|}, \sec x, e^x \right)$

(c) $\lim_{x \rightarrow 4} \left(\frac{x^2-16}{x+4}, x+7, \frac{\tan 4x}{5x} \right)$

(d) $\lim_{x \rightarrow \infty} \left(\frac{x}{1+x^2}, \frac{x^2}{1+x^2}, \frac{\sin x^2}{x} \right)$

2. Find

$$\lim_{x \rightarrow 2} \left(\frac{x^2-4}{x+2}, x^2+2x-1, \frac{x^2-4}{x-2} \right).$$

3. Prove from the definition that $\lim_{x \rightarrow a} (\sqrt[3]{x}, x+1) = (\sqrt[3]{a}, a+1)$ for all $a \in \mathbb{R}$. **Hint:** You might want to use the formula for the difference of two cubes,

$$a^3 - b^3 = (a-b)(a^2 + ab + b^2).$$

4. Let

$$\mathbf{r}(t) = (4 + t^2, \sqrt{t^2 + 1}t^3, t^3)$$

describe the position of an object in \mathbb{R}^3 as a function of t where t is measured in seconds and $\mathbf{r}(t)$ is measured in meters. Is the velocity of this object ever equal to zero? If so, find the value of t at which this occurs and the point in \mathbb{R}^3 at which the velocity is zero.

5. Let $\mathbf{r}(t) = (\sin 2t, t^2, 2t + 1)$ for $t \in [0, 4]$. Find a tangent line to the curve parameterized by \mathbf{r} at the point $\mathbf{r}(2)$.
6. Let $\mathbf{r}(t) = (t, \sin t^2, t + 1)$ for $t \in [0, 5]$. Find a tangent line to the curve parameterized by \mathbf{r} at the point $\mathbf{r}(2)$.
7. Let $\mathbf{r}(t) = (\sin t, t^2, \cos(t^2))$ for $t \in [0, 5]$. Find a tangent line to the curve parameterized by \mathbf{r} at the point $\mathbf{r}(2)$.
8. Let $\mathbf{r}(t) = (\sin t, \cos(t^2), t + 1)$ for $t \in [0, 5]$. Find the velocity when $t = 3$.
9. Let $\mathbf{r}(t) = (\sin t, t^2, t + 1)$ for $t \in [0, 5]$. Find the velocity when $t = 3$.
10. Let $\mathbf{r}(t) = (t, \ln(t^2 + 1), t + 1)$ for $t \in [0, 5]$. Find the velocity when $t = 3$.
11. Suppose an object has position $\mathbf{r}(t) \in \mathbb{R}^3$ where \mathbf{r} is differentiable and suppose also that $|\mathbf{r}(t)| = c$ where c is a constant.
 - (a) Show first that this condition does not require $\mathbf{r}(t)$ to be a constant. **Hint:** You can do this either mathematically or by giving a physical example.
 - (b) Show that you can conclude that $\mathbf{r}'(t) \cdot \mathbf{r}(t) = 0$. That is, the velocity is always perpendicular to the displacement.
12. Prove 16.4 from the component description of the cross product.
13. Prove 16.4 from the formula $(\mathbf{f} \times \mathbf{g})_i = \varepsilon_{ijk} f_j g_k$.
14. Prove 16.4 directly from the definition of the derivative without considering components.
15. A Bezier curve in \mathbb{R}^p is a vector valued function of the form

$$\mathbf{y}(t) = \sum_{k=0}^n \binom{n}{k} \mathbf{x}_k (1-t)^{n-k} t^k$$

where here the $\binom{n}{k}$ are the binomial coefficients and \mathbf{x}_k are $n + 1$ points in \mathbb{R}^n . Show that $\mathbf{y}(0) = \mathbf{x}_0$, $\mathbf{y}(1) = \mathbf{x}_n$, and find $\mathbf{y}'(0)$ and $\mathbf{y}'(1)$. Recall that $\binom{n}{0} = \binom{n}{n} = 1$ and $\binom{n}{n-1} = \binom{n}{1} = n$. Curves of this sort are important in various computer programs.

16. Suppose $\mathbf{r}(t)$, $\mathbf{s}(t)$, and $\mathbf{p}(t)$ are three differentiable functions of t which have values in \mathbb{R}^3 . Find a formula for $(\mathbf{r}(t) \times \mathbf{s}(t) \cdot \mathbf{p}(t))'$.
17. If $\mathbf{F}'(t) = \mathbf{f}(t)$ for all $t \in (a, b)$ and \mathbf{F} is continuous on $[a, b]$, show that $\int_a^b \mathbf{f}(t) dt = \mathbf{F}(b) - \mathbf{F}(a)$.

18. If $\mathbf{r}'(t) = \mathbf{0}$ for all $t \in (a, b)$, show that there exists a constant vector \mathbf{c} such that $\mathbf{r}(t) = \mathbf{c}$ for all $t \in (a, b)$.
19. Let $\mathbf{r}(t) = \left(\ln(t), \frac{t^2}{2}, \sqrt{2}t\right)$ for $t \in [1, 2]$. Find the length of this curve.
20. Let $\mathbf{r}(t) = \left(\frac{2}{3}t^{3/2}, t, t\right)$ for $t \in [0, 1]$. Find the length of this curve.
21. Let $\mathbf{r}(t) = (t, \cos(3t), \sin(3t))$ for $t \in [0, 1]$. Find the length of this curve.
22. Recall $\mathbf{p}'(t) = \lim_{h \rightarrow 0} \frac{\mathbf{p}(t+h) - \mathbf{p}(t)}{h}$. Show that this is equivalent to saying either of the following.

$$\begin{aligned}\mathbf{p}(t+h) - \mathbf{p}(t) &= \mathbf{p}'(t)h + o(h) \\ \mathbf{p}(t) - \mathbf{p}(s) &= \mathbf{p}'(s)(t-s) + o(t-s)\end{aligned}$$

where $\lim_{h \rightarrow 0} \frac{o(h)}{h} = \mathbf{0}$.

23. Recall that the length of a curve is approximated by the length of a polygonal curve $\sum_{k=1}^m |\mathbf{p}(t_k) - \mathbf{p}(t_{k-1})|$ where $a = t_0 < \cdots < t_m = b$. Letting the norm of the partition $P = \{t_0, \dots, t_m\}$ be small enough, argue that from differentiability,

$$\begin{aligned}\sum_{k=1}^m |\mathbf{p}'(t_{k-1})| (t_k - t_{k-1}) - \varepsilon(b-a) \\ \sum_{k=1}^m |\mathbf{p}(t_k) - \mathbf{p}(t_{k-1})| \leq \sum_{k=1}^m |\mathbf{p}'(t_{k-1})| (t_k - t_{k-1}) + \varepsilon(b-a)\end{aligned}$$

Explain why if you let $\|P_k\| \rightarrow 0$,

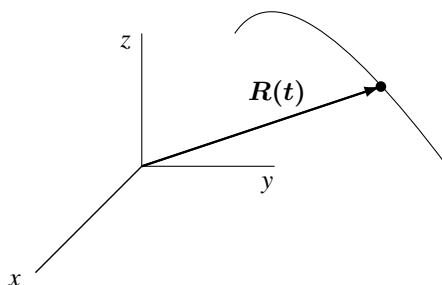
$$\begin{aligned}\limsup_{k \rightarrow \infty} \left(\sum_{t_k \in P_k} |\mathbf{p}(t_k) - \mathbf{p}(t_{k-1})| - \int_a^b |\mathbf{p}'(t)| dt \right) \\ - \liminf_{k \rightarrow \infty} \left(\sum_{t_k \in P_k} |\mathbf{p}(t_k) - \mathbf{p}(t_{k-1})| - \int_a^b |\mathbf{p}'(t)| dt \right) \\ \leq 2\varepsilon(b-a)\end{aligned}$$

Now explain why $\lim_{k \rightarrow \infty} \sum_{t_k \in P_k} |\mathbf{p}(t_k) - \mathbf{p}(t_{k-1})| = \int_a^b |\mathbf{p}'(t)| dt$. This gives a proof that the lengths of polygonal curves approximating the curve converge to the appropriate integral. Thus we could have defined the length as the limit of the lengths of the approximating polygonal curves and replaced the definition with a theorem.

16.6 Motion on Space Curves

A fly buzzing around the room, a person riding a roller coaster, and a satellite orbiting the earth all have something in common. They are moving over some sort of curve in three dimensions.

Denote by $\mathbf{R}(t)$ the position vector of the point on the curve which occurs at time t . Assume that $\mathbf{R}', \mathbf{R}''$ exist and are continuous. Thus $\mathbf{R}' = \mathbf{v}$, the velocity and $\mathbf{R}'' = \mathbf{a}$ is defined as the acceleration.



Lemma 16.6.1 Define $\mathbf{T}(t) \equiv \mathbf{R}'(t) / |\mathbf{R}'(t)|$. Then $|\mathbf{T}(t)| = 1$ and if $\mathbf{T}'(t) \neq 0$, then there exists a unit vector $\mathbf{N}(t)$ perpendicular to $\mathbf{T}(t)$ and a scalar valued function $\kappa(t)$, with $\mathbf{T}'(t) = \kappa(t) |\mathbf{v}| \mathbf{N}(t)$.

Proof: It follows from the definition that $|\mathbf{T}| = 1$. Therefore, $\mathbf{T} \cdot \mathbf{T} = 1$ and so, upon differentiating both sides, $\mathbf{T}' \cdot \mathbf{T} + \mathbf{T} \cdot \mathbf{T}' = 2\mathbf{T}' \cdot \mathbf{T} = 0$. Therefore, \mathbf{T}' is perpendicular to \mathbf{T} . Let $\mathbf{N}(t) |\mathbf{T}'| \equiv \mathbf{T}'$. Note that if $|\mathbf{T}'| = 0$, you could let $\mathbf{N}(t)$ be any unit vector. Then letting $\kappa(t)$ be defined such that $|\mathbf{T}'| \equiv \kappa(t) |\mathbf{v}|$, it follows $\mathbf{T}'(t) = |\mathbf{T}'(t)| \mathbf{N}(t) = \kappa(t) |\mathbf{v}| \mathbf{N}(t)$. ■

Definition 16.6.2 The vector $\mathbf{T}(t)$ is called the **unit tangent vector** and the vector $\mathbf{N}(t)$ is called the **principal normal**. The function $\kappa(t)$ in the above lemma is called the **curvature**. The **radius of curvature** is defined as $\rho = 1/\kappa$. The plane determined by the two vectors \mathbf{T} and \mathbf{N} in the case where $\mathbf{T}' \neq 0$ is called the **osculating¹ plane**. It identifies a particular plane which is in a sense tangent to this space curve.

The important thing about this is that it is possible to write the acceleration as the sum of two vectors, one perpendicular to the direction of motion and the other in the direction of motion.

Theorem 16.6.3 For $\mathbf{R}(t)$ the position vector of a space curve, the acceleration is given by the formula

$$\mathbf{a} = \frac{d|\mathbf{v}|}{dt} \mathbf{T} + \kappa |\mathbf{v}|^2 \mathbf{N} \equiv a_T \mathbf{T} + a_N \mathbf{N}. \quad (16.11)$$

Furthermore, $a_T^2 + a_N^2 = |\mathbf{a}|^2$.

Proof: $\mathbf{a} = \frac{d\mathbf{v}}{dt} = \frac{d}{dt} (\mathbf{R}') = \frac{d}{dt} (|\mathbf{v}| \mathbf{T}) = \frac{d|\mathbf{v}|}{dt} \mathbf{T} + |\mathbf{v}| \mathbf{T}' = \frac{d|\mathbf{v}|}{dt} \mathbf{T} + |\mathbf{v}|^2 \kappa \mathbf{N}$. This proves the first part.

For the second part,

$$\begin{aligned} |\mathbf{a}|^2 &= (a_T \mathbf{T} + a_N \mathbf{N}) \cdot (a_T \mathbf{T} + a_N \mathbf{N}) \\ &= a_T^2 \mathbf{T} \cdot \mathbf{T} + 2a_N a_T \mathbf{T} \cdot \mathbf{N} + a_N^2 \mathbf{N} \cdot \mathbf{N} = a_T^2 + a_N^2 \end{aligned}$$

because $\mathbf{T} \cdot \mathbf{N} = 0$. ■

From 16.11 and the geometric properties of the cross product,

$$\mathbf{a} \times \mathbf{v} = \kappa |\mathbf{v}|^2 \mathbf{N} \times \mathbf{v}$$

¹To osculate means to kiss. Thus this plane could be called the kissing plane. However, that does not sound formal enough so we call it the osculating plane.

Hence, using the geometric description of the cross product again using that the angle between \mathbf{N} and \mathbf{T} is 90° ,

$$|\mathbf{a} \times \mathbf{v}| = \kappa |\mathbf{v}|^2 |\mathbf{v}|, \quad \kappa = \frac{|\mathbf{a} \times \mathbf{v}|}{|\mathbf{v}|^3} = \frac{|\mathbf{v} \times \mathbf{a}|}{|\mathbf{v}|^3} \quad (16.12)$$

Finally, it is good to point out that the curvature is a property of the curve itself, and does not depend on the parametrization of the curve. If the curve is given by two different vector valued functions $\mathbf{R}(t)$ and $\mathbf{R}(\tau)$, then from the formula above for the curvature,

$$\kappa(t) = \frac{|\mathbf{T}'(t)|}{|\mathbf{v}(t)|} = \frac{\left| \frac{d\mathbf{T}}{d\tau} \frac{d\tau}{dt} \right|}{\left| \frac{d\mathbf{R}}{d\tau} \frac{d\tau}{dt} \right|} = \frac{\left| \frac{d\mathbf{T}}{d\tau} \right|}{\left| \frac{d\mathbf{R}}{d\tau} \right|} \equiv \kappa(\tau).$$

From this, it is possible to give an important formula from physics. Suppose an object orbits a point at constant speed v and it travels over a circle of radius r . In the above notation, $|\mathbf{v}| = v$. What is the centripetal acceleration of this object? You may know from a physics class that the answer is v^2/r where r is the radius. This follows from the above quite easily. First, what is the curvature of a circle of radius r ? A parameterization of such a curve is $\mathbf{R}(t) = (r \cos t, r \sin t)$. Thus using 16.12 and this parametrization,

$$\mathbf{v} \times \mathbf{a} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ -r \sin t & r \cos t & 0 \\ -r \cos t & -r \sin t & 0 \end{vmatrix} = kr^2$$

Thus $\kappa = \frac{r^2}{r^3} = \frac{1}{r}$. Since v is constant, it follows from 16.11 that

$$\mathbf{a} = \frac{1}{r} |\mathbf{v}|^2 \mathbf{N} = \frac{1}{r} v^2 \mathbf{N}$$

16.6.1 Some Simple Techniques

Recall the formula for acceleration is

$$\mathbf{a} = a_T \mathbf{T} + a_N \mathbf{N} \quad (16.13)$$

where $a_T = \frac{d|\mathbf{v}|}{dt}$ and $a_N = \kappa |\mathbf{v}|^2$. Of course one way to find a_T and a_N is to just find $|\mathbf{v}|$, $\frac{d|\mathbf{v}|}{dt}$ and κ and plug in. However, there is another way which might be easier. Take the dot product of both sides with \mathbf{T} a vector which is easy to find. This gives,

$$\mathbf{a} \cdot \mathbf{T} = a_T \mathbf{T} \cdot \mathbf{T} + a_N \mathbf{N} \cdot \mathbf{T} = a_T.$$

Thus

$$\mathbf{a} = (\mathbf{a} \cdot \mathbf{T}) \mathbf{T} + a_N \mathbf{N}$$

and so

$$\mathbf{a} - (\mathbf{a} \cdot \mathbf{T}) \mathbf{T} = a_N \mathbf{N} \quad (16.14)$$

and taking norms of both sides,

$$|\mathbf{a} - (\mathbf{a} \cdot \mathbf{T}) \mathbf{T}| = a_N.$$

Also from 16.14,

$$\frac{\mathbf{a} - (\mathbf{a} \cdot \mathbf{T})\mathbf{T}}{|\mathbf{a} - (\mathbf{a} \cdot \mathbf{T})\mathbf{T}|} = \frac{a_N \mathbf{N}}{a_N |\mathbf{N}|} = \mathbf{N}.$$

Also recall

$$\kappa = \frac{|\mathbf{a} \times \mathbf{v}|}{|\mathbf{v}|^3}, \quad a_T^2 + a_N^2 = |\mathbf{a}|^2$$

This is usually easier than computing $\mathbf{T}'/|\mathbf{T}'|$. To illustrate the use of these simple observations, here is a simple example.

Example 16.6.4 Let $\mathbf{R}(t) = (\cos(t), t, t^2)$ for $t \in [0, 3]$. Find the speed, velocity, curvature, and write the acceleration in terms of normal and tangential components when $t = 0$. Also find \mathbf{N} at the point where $t = 0$.

First I need to find the velocity and acceleration. Thus

$$\mathbf{v} = (-\sin t, 1, 2t), \quad \mathbf{a} = (-\cos t, 0, 2)$$

and consequently, $\mathbf{T} = \frac{(-\sin t, 1, 2t)}{\sqrt{\sin^2(t) + 1 + 4t^2}}$. When $t = 0$, this reduces to

$$\mathbf{v}(0) = (0, 1, 0), \quad \mathbf{a} = (-1, 0, 2), \quad |\mathbf{v}(0)| = 1, \quad \mathbf{T} = (0, 1, 0).$$

Then the tangential component of acceleration when $t = 0$ is

$$a_T = (-1, 0, 2) \cdot (0, 1, 0) = 0$$

Now $|\mathbf{a}|^2 = 5$ and so $a_N = \sqrt{5}$ because $a_T^2 + a_N^2 = |\mathbf{a}|^2$. Thus $\sqrt{5} = \kappa |\mathbf{v}(0)|^2 = \kappa \cdot 1 = \kappa$. Next let's find \mathbf{N} . From $\mathbf{a} = a_T \mathbf{T} + a_N \mathbf{N}$ it follows

$$(-1, 0, 2) = 0 \cdot \mathbf{T} + \sqrt{5} \mathbf{N}$$

and so

$$\mathbf{N} = \frac{1}{\sqrt{5}}(-1, 0, 2).$$

This was pretty easy.

Example 16.6.5 Find a formula for the curvature of the curve given by the graph of $y = f(x)$ for $x \in [a, b]$. Assume whatever you like about smoothness of f .

You need to write this as a parametric curve. This is most easily accomplished by letting $t = x$. Thus a parametrization is $(t, f(t), 0) : t \in [a, b]$. Then you can use the formula given above. The acceleration is $(0, f''(t), 0)$ and the velocity is $(1, f'(t), 0)$. Therefore,

$$\mathbf{a} \times \mathbf{v} = (0, f''(t), 0) \times (1, f'(t), 0) = (0, 0, -f''(t)).$$

Therefore, the curvature is given by

$$\frac{|\mathbf{a} \times \mathbf{v}|}{|\mathbf{v}|^3} = \frac{|f''(t)|}{(1 + f'(t)^2)^{3/2}}.$$

Sometimes curves do not come to you parametrically. This is unfortunate when it occurs but you can sometimes find a parametric description of such curves. It should be emphasized that it is only sometimes when you can actually find a parametrization. General systems of nonlinear equations cannot be solved using algebra.

Example 16.6.6 Find a parametrization for the intersection of the surfaces

$$y + 3z = 2x^2 + 4 \text{ and } y + 2z = x + 1.$$

You need to solve for x and y in terms of x . This yields

$$z = 2x^2 - x + 3, \quad y = -4x^2 + 3x - 5.$$

Therefore, letting $t = x$, the parametrization is

$$(x, y, z) = (t, -4t^2 - 5 + 3t, -t + 3 + 2t^2).$$

Example 16.6.7 Find a parametrization for the straight line joining $(3, 2, 4)$ and $(1, 10, 5)$.

$(x, y, z) = (3, 2, 4) + t(-2, 8, 1) = (3 - 2t, 2 + 8t, 4 + t)$ where $t \in [0, 1]$. Note where this came from. The vector $(-2, 8, 1)$ is obtained from $(1, 10, 5) - (3, 2, 4)$. Now you should check to see this works. It is usually not possible to find an explicit formula for the intersection of two surfaces as was just done.

16.7 Geometry of Space Curves*

If you are interested in more on space curves, you should read this section. Otherwise, proceed to the exercises. Denote by $\mathbf{R}(s)$ the function which takes s to a point on this curve where s is arc length. Thus $\mathbf{R}(s)$ equals the point on the curve which occurs when you have traveled a distance of s along the curve from one end. This is known as the parametrization of the curve in terms of arc length. Note also that it incorporates an orientation on the curve because there are exactly two ends you could begin measuring length from. In this section, assume anything about smoothness and continuity to make the following manipulations valid. In particular, assume that \mathbf{R}' exists and is continuous.

Lemma 16.7.1 Define $\mathbf{T}(s) \equiv \mathbf{R}'(s)$. Then $|\mathbf{T}(s)| = 1$ and if $\mathbf{T}'(s) \neq 0$, then there exists a unit vector $\mathbf{N}(s)$ perpendicular to $\mathbf{T}(s)$ and a scalar valued function $\kappa(s)$ with $\mathbf{T}'(s) = \kappa(s) \mathbf{N}(s)$.

Proof: First, $s = \int_0^s |\mathbf{R}'(r)| dr$ because of the definition of arc length. Therefore, from the fundamental theorem of calculus, $1 = |\mathbf{R}'(s)| = |\mathbf{T}(s)|$. Therefore, $\mathbf{T} \cdot \mathbf{T} = 1$ and so upon differentiating this on both sides, yields $\mathbf{T}' \cdot \mathbf{T} + \mathbf{T} \cdot \mathbf{T}' = 0$ which shows $\mathbf{T} \cdot \mathbf{T}' = 0$. Therefore, the vector \mathbf{T}' is perpendicular to the vector \mathbf{T} . In case $\mathbf{T}'(s) \neq 0$, let $\mathbf{N}(s) = \frac{\mathbf{T}'(s)}{|\mathbf{T}'(s)|}$ and so $\mathbf{T}'(s) = |\mathbf{T}'(s)| \mathbf{N}(s)$, showing the scalar valued function is $\kappa(s) = |\mathbf{T}'(s)|$. ■

The radius of curvature is defined as $\rho = \frac{1}{\kappa}$. Thus at points where there is a lot of curvature, the radius of curvature is small and at points where the curvature is small, the radius of curvature is large. The plane determined by the two vectors \mathbf{T} and \mathbf{N} is called the osculating plane. It identifies a particular plane which is in a sense tangent to this space curve. In the case where $|\mathbf{T}'(s)| = 0$ near the point of interest, $\mathbf{T}(s)$ equals a constant and so the space curve is a straight line which it would be supposed has no curvature. Also, the principal normal is undefined in this case. This makes sense because if there is no curving going on, there is no special direction normal to the curve at such points which could be distinguished from any other direction normal to the curve. In the case where $|\mathbf{T}'(s)| = 0$, $\kappa(s) = 0$ and the radius of curvature would be considered infinite.

Definition 16.7.2 The vector $T(s)$ is called the unit tangent vector and the vector $N(s)$ is called the **principal normal**. The function $\kappa(s)$ in the above lemma is called the **curvature**. When $T'(s) \neq 0$ so the principal normal is defined, the vector $B(s) \equiv T(s) \times N(s)$ is called the **binormal**.

The binormal is normal to the osculating plane and B' tells how fast this vector changes. Thus it measures the rate at which the curve twists.

Lemma 16.7.3 Let $R(s)$ be a parametrization of a space curve with respect to arc length and let the vectors T, N , and B be as defined above. Then $B' = T \times N'$ and there exists a scalar function $\tau(s)$ such that $B' = \tau N$.

Proof: From the definition of $B = T \times N$, and you can differentiate both sides and get $B' = T' \times N + T \times N'$. Now recall that T' is a multiple called curvature multiplied by N so the vectors T' and N have the same direction, so $B' = T \times N'$. Therefore, B' is either zero or is perpendicular to T . But also, from the definition of B , B is a unit vector and so $B(s) \cdot B(s) = 1$. Differentiating this, $B'(s) \cdot B(s) + B(s) \cdot B'(s) = 0$ showing that B' is perpendicular to B also. Therefore, B' is a vector which is perpendicular to both vectors T and B and since this is in three dimensions, B' must be some scalar multiple of N , and this multiple is called τ . Thus $B' = \tau N$ as claimed. ■

Lets go over this last claim a little more. The following situation is obtained. There are two vectors T and B which are perpendicular to each other and both B' and N are perpendicular to these two vectors, hence perpendicular to the plane determined by them. Therefore, B' must be a multiple of N . Take a piece of paper, draw two unit vectors on it which are perpendicular. Then you can see that any two vectors which are perpendicular to this plane must be multiples of each other.

The scalar function τ is called the torsion. In case $T' = 0$, none of this is defined because in this case there is not a well defined osculating plane. The conclusion of the following theorem is called the Serret Frenet formulas.

Theorem 16.7.4 (Serret Frenet) Let $R(s)$ be the parametrization with respect to arc length of a space curve and $T(s) = R'(s)$ is the unit tangent vector. Suppose $|T'(s)| \neq 0$ so the principal normal $N(s) = \frac{T'(s)}{|T'(s)|}$ is defined. The binormal is the vector $B \equiv T \times N$ so T, N, B forms a right handed system of unit vectors each of which is perpendicular to every other. Then the following system of differential equations holds in \mathbb{R}^9 .

$$B' = \tau N, T' = \kappa N, N' = -\kappa T - \tau B$$

where κ is the curvature and is nonnegative and τ is the **torsion**.

Proof: $\kappa \geq 0$ because $\kappa = |T'(s)|$. The first two equations are already established. To get the third, note that $B \times T = N$ which follows because T, N, B is given to form a right handed system of unit vectors each perpendicular to the others. (Use your right hand.) Now take the derivative of this expression. thus

$$N' = B' \times T + B \times T' = \tau N \times T + \kappa B \times N.$$

Now recall again that T, N, B is a right hand system. Thus

$$N \times T = -B, B \times N = -T.$$

This establishes the Frenet Serret formulas. ■

This is an important example of a system of differential equations in \mathbb{R}^9 . It is a remarkable result because it says that from knowledge of the two scalar functions τ and κ , and initial values for \mathbf{B} , \mathbf{T} , and \mathbf{N} when $s = 0$ you can obtain the binormal, unit tangent, and principal normal vectors. It is just the solution of an initial value problem although this is for a vector valued rather than scalar valued function. Having done this, you can reconstruct the entire space curve starting at some point \mathbf{R}_0 because $\mathbf{R}'(s) = \mathbf{T}(s)$ and so $\mathbf{R}(s) = \mathbf{R}_0 + \int_0^s \mathbf{T}(r) dr$. There are ways to solve such a system of equations numerically and even draw the graph of the resulting curve but this is not a topic for this book.

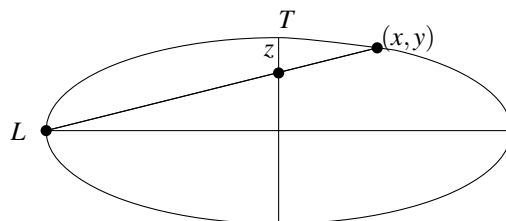
16.8 Exercises

- Find a parametrization for the intersection of the planes $2x + y + 3z = -2$ and $3x - 2y + z = -4$.
- Find a parametrization for the intersection of the plane $3x + y + z = -3$ and the circular cylinder $x^2 + y^2 = 1$.
- Find a parametrization for the intersection of the plane $4x + 2y + 3z = 2$ and the elliptic cylinder $x^2 + 4z^2 = 9$.
- Find a parametrization for the straight line joining $(1, 2, 1)$ and $(-1, 4, 4)$.
- Find a parametrization for the intersection of the surfaces $3y + 3z = 3x^2 + 2$ and $3y + 2z = 3$.
- Find a formula for the curvature of the curve $y = \sin x$ in the xy plane.
- An object moves over the curve (t, e^t, at) where $t \in \mathbb{R}$ and a is a positive constant. Find the value of t at which the normal component of acceleration is largest if there is such a point.
- Find a formula for the curvature of the space curve in \mathbb{R}^2 , $(x(t), y(t))$.
- An object moves over the helix, $(\cos 3t, \sin 3t, 5t)$. Find the normal and tangential components of the acceleration of this object as a function of t and write the acceleration in the form $a_T \mathbf{T} + a_N \mathbf{N}$.
- An object moves in \mathbb{R}^3 according to the formula $(\cos 3t, \sin 3t, t^2)$. Find the normal and tangential components of the acceleration of this object as a function of t and write the acceleration in the form $a_T \mathbf{T} + a_N \mathbf{N}$.
- An object moves over the helix, $(\cos t, \sin t, 2t)$. Find the osculating plane at the point of the curve corresponding to $t = \pi/4$.
- An object moves over a circle of radius r according to the formula

$$\mathbf{r}(t) = (r \cos(\omega t), r \sin(\omega t))$$

where $v = r\omega$. Show that the speed of the object is constant and equals to v . Tell why $a_T = 0$ and find a_N , \mathbf{N} .

13. Suppose $|\mathbf{R}(t)| = c$ where c is a constant. Show the velocity, $\mathbf{R}'(t)$ is always perpendicular to $\mathbf{R}(t)$.
14. An object moves in three dimensions and the only force on the object is a central force. This means that if $\mathbf{r}(t)$ is the position of the object, $\mathbf{a}(t) = k(\mathbf{r}(t))\mathbf{r}(t)$ where k is some function. Show that if this happens, then the motion of the object must be in a plane. **Hint:** First argue that $\mathbf{a} \times \mathbf{r} = \mathbf{0}$. Next show that $(\mathbf{a} \times \mathbf{r})' = (\mathbf{v} \times \mathbf{r})'$. Therefore, $(\mathbf{v} \times \mathbf{r})' = \mathbf{0}$. Explain why this requires $\mathbf{v} \times \mathbf{r} = \mathbf{c}$ for some vector \mathbf{c} which does not depend on t . Then explain why $\mathbf{c} \cdot \mathbf{r} = 0$. This implies the motion is in a plane. Why? What are some examples of central forces?
15. Let $\mathbf{R}(t) = (\cos t)\mathbf{i} + (\cos t)\mathbf{j} + (\sqrt{2}\sin t)\mathbf{k}$. Find the arc length, s as a function of the parameter t , if $t = 0$ is taken to correspond to $s = 0$.
16. Let $\mathbf{R}(t) = 2\mathbf{i} + (4t + 2)\mathbf{j} + 4t\mathbf{k}$. Find the arc length, s as a function of the parameter t , if $t = 0$ is taken to correspond to $s = 0$.
17. Let $\mathbf{R}(t) = e^{5t}\mathbf{i} + e^{-5t}\mathbf{j} + 5\sqrt{2}t\mathbf{k}$. Find the arc length, s as a function of the parameter t , if $t = 0$ is taken to correspond to $s = 0$.
18. Consider the curve obtained from the graph of $y = f(x)$. Find a formula for the curvature.
19. Consider the curve in the plane $y = e^x$. Find the point on this curve at which the curvature is a maximum.
20. An object moves along the x axis toward $(0,0)$ and then along the curve $y = x^2$ in the direction of increasing x at constant speed. Is the force acting on the object a continuous function? Explain. Is there any physically reasonable way to make this force continuous by relaxing the requirement that the object move at constant speed? If the curve were part of a railroad track, what would happen at the point where $x = 0$?
21. An object of mass m moving over a space curve is acted on by a force, \mathbf{F} . The work is defined as $\int_a^b \mathbf{F}(\mathbf{r}(t)) \cdot \mathbf{r}'(t) dt$ and recall that $\mathbf{F} = m\mathbf{a}$, the mass times the acceleration. Work will be discussed more formally later. Show the work done by this force equals ma_T (length of the curve). In other words, it is only the tangential component of the force which does work.
22. The edge of an elliptical skating rink represented in the following picture has a light at its left end and satisfies the equation $\frac{x^2}{900} + \frac{y^2}{256} = 1$. (Distances measured in yards.)



A hockey puck slides from the point T towards the center of the rink at the rate of 2 yards per second. What is the speed of its shadow along the wall when $z = 8$? **Hint:** You need to find $\sqrt{x'^2 + y'^2}$ at the instant described.

23. Use MATLAB to graph the parametric curve $x = t \cos(t)$, $y = t \sin(t)$, $z = t^2$ for $t \in [0, 24]$.

Chapter 17

Some Physical Applications

17.1 Spherical and Cylindrical Coordinates

There are two extensions of polar coordinates to three dimensions which are important in applications, cylindrical and spherical coordinates. These will be studied much more in multi-variable calculus but it is convenient to give an introduction to these here. It is important to understand the geometric significance of these coordinate systems. When you remember the geometric meaning of the spherical coordinates, they are not too bad, but if you try to ignore this, you will be constantly confused about what you are trying to do.

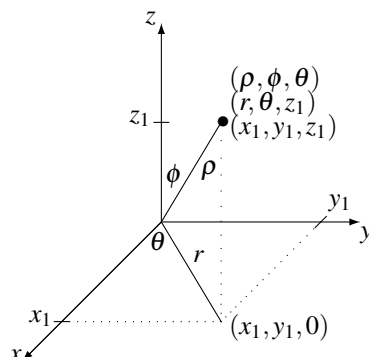
Cylindrical coordinates are defined as follows.

$$\begin{aligned} \mathbf{x}(r, \theta, z) &\equiv \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} r \cos(\theta) \\ r \sin(\theta) \\ z \end{pmatrix}, \\ r &\geq 0, \theta \in [0, 2\pi), z \in \mathbb{R} \end{aligned}$$

Spherical coordinates are a little harder. These are given by

$$\begin{aligned} \mathbf{x}(\rho, \theta, \phi) &\equiv \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \rho \sin(\phi) \cos(\theta) \\ \rho \sin(\phi) \sin(\theta) \\ \rho \cos(\phi) \end{pmatrix}, \\ \rho &\geq 0, \theta \in [0, 2\pi), \phi \in [0, \pi] \end{aligned}$$

The following picture relates the various coordinates.



In this picture, ρ is the distance between the origin, the point whose Cartesian coordinates are $(0,0,0)$ and the point indicated by a dot and labelled as (x_1, y_1, z_1) , (r, θ, z_1) , and (ρ, ϕ, θ) . The angle between the positive z axis and the line between the origin and the point indicated by a dot is denoted by ϕ , and θ is the angle between the positive x axis and the line joining the origin to the point $(x_1, y_1, 0)$ as shown, while r is the length of this line. Thus $r = \rho \sin(\phi)$ and is the usual polar coordinate while θ is the other polar coordinate. Letting z_1 denote the usual z coordinate of a point in three dimensions, like the one shown as a dot, (r, θ, z_1) are the cylindrical coordinates of the dotted point. The spherical coordinates are determined by (ρ, ϕ, θ) . When ρ is specified, this indicates that the point of interest is on some sphere of radius ρ which is centered at the origin. Then when ϕ is given, the location of the point is narrowed down to a circle of “latitude” and finally, θ determines which point is on this circle by specifying a circle of “longitude”. Let $\phi \in [0, \pi]$, $\theta \in [0, 2\pi)$, and $\rho \in [0, \infty)$. The picture shows how to relate these new coordinate systems to Cartesian coordinates. Note that θ is the same in the two coordinate systems and that $\rho \sin \phi = r$.

If you fix two of the variables and take a derivative with respect to the other, you are finding the tangent vector to a space curve. There are three of these space curves

$$\rho \rightarrow \begin{pmatrix} \rho \sin(\phi) \cos(\theta) \\ \rho \sin(\phi) \sin(\theta) \\ \rho \cos(\phi) \end{pmatrix}, \theta \rightarrow \begin{pmatrix} \rho \sin(\phi) \cos(\theta) \\ \rho \sin(\phi) \sin(\theta) \\ \rho \cos(\phi) \end{pmatrix}, \phi \rightarrow \begin{pmatrix} \rho \sin(\phi) \cos(\theta) \\ \rho \sin(\phi) \sin(\theta) \\ \rho \cos(\phi) \end{pmatrix}$$

Doing derivatives with respect to ρ, θ , and ϕ for these three space curves gives tangent vectors, the first in the direction of increasing ρ for fixed θ, ϕ , the second in the direction of increasing θ fixing ρ, ϕ , and the third in the direction of increasing ϕ for fixed θ, ρ . These tangent vectors are

$$\begin{pmatrix} \sin(\phi) \cos(\theta) \\ \sin(\phi) \sin(\theta) \\ \cos(\phi) \end{pmatrix}, \begin{pmatrix} -\rho \sin \theta \sin \phi \\ \rho \cos \theta \sin \phi \\ 0 \end{pmatrix}, \begin{pmatrix} \rho \cos \theta \cos \phi \\ \rho \cos \phi \sin \theta \\ -\rho \sin \phi \end{pmatrix}$$

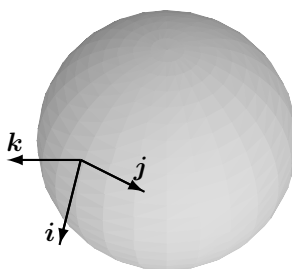
You should note that the dot product of any two different vectors is zero. This is why these spherical coordinates are known as an orthogonal. This procedure is important for general curvilinear coordinates.

It is often convenient to divide these vectors by their lengths to obtain unit vectors in the given directions. Also it is very useful to list them in an order that the vectors are a right handed system. I will do this later by listing them according to differentiating with respect

to ϕ first, then with respect to θ and then with respect to ρ . This yields the following unit vectors in this order:

$$\begin{pmatrix} \cos \theta \cos \phi \\ \cos \phi \sin \theta \\ -\sin \phi \end{pmatrix}, \begin{pmatrix} -\sin \theta \\ \cos \theta \\ 0 \end{pmatrix}, \begin{pmatrix} \sin(\phi) \cos(\theta) \\ \sin(\phi) \sin(\theta) \\ \cos(\phi) \end{pmatrix}$$

You could denote these vectors respectively as i, j, k and from the geometrical definition of the cross product, it follows that $i \times j = k$, $j \times k = i$, etc. Here is a picture illustrating these vectors in the order just described at a point (ρ, θ, ϕ) of the sphere of radius ρ . The first is tangent to a line of longitude, the second, a line of latitude and the third points directly out away from the sphere of radius ρ .



17.2 Exercises

- The following are the polar coordinates of points. Find the rectangular coordinates.

- $(5, \frac{\pi}{6})$
- $(3, \frac{\pi}{3})$
- $(4, \frac{2\pi}{3})$
- $(2, \frac{3\pi}{4})$
- $(3, \frac{7\pi}{6})$
- $(8, \frac{11\pi}{6})$

- The following are the rectangular coordinates of points. Find the polar coordinates of these points.

- $(\frac{5}{2}\sqrt{2}, \frac{5}{2}\sqrt{2})$
- $(\frac{3}{2}, \frac{3}{2}\sqrt{3})$
- $(-\frac{5}{2}\sqrt{2}, \frac{5}{2}\sqrt{2})$
- $(-\frac{5}{2}, \frac{5}{2}\sqrt{3})$

(e) $(-\sqrt{3}, -1)$

(f) $(\frac{3}{2}, -\frac{3}{2}\sqrt{3})$

3. The spherical coordinates are given. Find the rectangular coordinates (x, y, z) . (It is typically a nuisance to go the other direction and in practice, you don't want to do this anyway.)

(a) $(\rho, \theta, \phi) = (4, \frac{\pi}{2}, \frac{\pi}{2})$

(b) $(\rho, \theta, \phi) = (1, \frac{\pi}{3}, \frac{2\pi}{3})$

(c) $(\rho, \theta, \phi) = (2, \frac{3}{2}\pi, \frac{\pi}{3})$

(d) $(\rho, \theta, \phi) = (3, \frac{2\pi}{3}, \frac{5\pi}{6})$

4. Find the tangent vectors corresponding to keeping two coordinates constant in the case of cylindrical coordinates and verify that cylindrical coordinates are orthogonal like spherical coordinates.

5. Verify that the vectors

$$\left(\begin{pmatrix} \cos \theta \cos \phi \\ \cos \phi \sin \theta \\ -\sin \phi \end{pmatrix}, \begin{pmatrix} -\sin \theta \\ \cos \theta \\ 0 \end{pmatrix}, \begin{pmatrix} \sin(\phi) \cos(\theta) \\ \sin(\phi) \sin(\theta) \\ \cos(\phi) \end{pmatrix} \right)$$

obtained from differentiating the spherical coordinates with respect to ϕ, θ , and ρ and then dividing by the length are an orthonormal right handed system of vectors.

6. In general it is a stupid idea to try to use algebra to invert and solve for a set of curvilinear coordinates such as polar or cylindrical coordinates in term of Cartesian coordinates. Not only is it often very difficult or even impossible to do it¹, but also it takes you in entirely the wrong direction because the whole point of introducing the new coordinates is to write everything in terms of these new coordinates and not in terms of Cartesian coordinates. However, sometimes this inversion can be done. Describe how to solve for r and θ in terms of x and y in polar coordinates.

17.3 Planetary Motion

Suppose at each point of space, \mathbf{r} is associated a force $\mathbf{F}(\mathbf{r})$ which a given object of mass m will experience if its position vector is \mathbf{r} . This is called a force field. a force field is a central force field if $\mathbf{F}(\mathbf{r}) = g(\mathbf{r})\mathbf{e}_r$. Thus in a central force field the force an object experiences will always be directed toward or away from the origin, $\mathbf{0}$. The following simple lemma is very interesting because it says that in a central force field objects must move in a plane.

Lemma 17.3.1 *Suppose an object moves in three dimensions in such a way that the only force acting on the object is a central force. Then the motion of the object is in a plane.*

¹It is no problem for these simple cases of curvilinear coordinates. However, it is a major difficulty in general. Algebra is simply not adequate to solve systems of nonlinear equations.

Proof: Let $\mathbf{r}(t)$ denote the position vector of the object. Then from the definition of a central force and Newton's second law,

$$m\mathbf{r}'' = g(\mathbf{r})\mathbf{r}.$$

Therefore,

$$m\mathbf{r}'' \times \mathbf{r} = m(\mathbf{r}' \times \mathbf{r})' = g(\mathbf{r})\mathbf{r} \times \mathbf{r} + m\mathbf{r}' \times \mathbf{r}' = \mathbf{0}.$$

Therefore, $(\mathbf{r}' \times \mathbf{r}) = \mathbf{n}$, a constant vector and so $\mathbf{r} \cdot \mathbf{n} = \mathbf{r} \cdot (\mathbf{r}' \times \mathbf{r}) = 0$ showing that \mathbf{n} is a normal vector to a plane which contains $\mathbf{r}(t)$ for all t . ■

Kepler's laws of planetary motion state, among other things, that planets move around the sun along an ellipse. These laws, discovered by Kepler, were shown by Newton to be consequences of his law of gravitation which states that the force acting on a mass m by a mass M is given by

$$\mathbf{F} = -GMm \left(\frac{1}{r^3} \right) \mathbf{r} = -GMm \left(\frac{1}{r^2} \right) \mathbf{e}_r$$

where r is the distance between centers of mass and \mathbf{r} is the position vector from M to m . Here G is the gravitation constant. This is called an inverse square law. Gravity acts according to this law and so does electrostatic force. The constant G , is very small when usual units are used and it has been computed using a very delicate experiment. It is now accepted to be

$$6.67 \times 10^{-11} \text{ Newton meter}^2/\text{kilogram}^2.$$

The experiment involved a light source shining on a mirror attached to a fiber from which was suspended a long rod with two solid balls of equal mass at the ends which were attracted by two larger masses. The gravitation force between the suspended balls and the two large balls caused the fibre to twist ever so slightly and this twisting was measured by observing the deflection of the light reflected from the mirror on a scale placed some distance from the fibre. Part of the experiment must compute the necessary spring constant of the fibre.

This constant was first measured successfully by Cavendish in 1798 in the manner just described. The accelerations are extremely small so it took months to complete the experiment. Also, the entire apparatus had to be shielded from any currents of air which would of course render the results worthless. The measurement has been made repeatedly. You should also note that it also depends on being able to show that the entire force can be considered as acting between the centers of mass of the respective balls. However, this was shown by Newton. If you have spherical coordinates which are curvilinear coordinates in three dimensions, this is not too hard, but none of this was invented in Newton's time.

In the following argument, M is the mass of the sun and m is the mass of the planet. (It could also be a comet or an asteroid.)

17.3.1 The Equal Area Rule, Kepler's Second Law

An object moves in three dimensions in such a way that the only force acting on the object is a central force. Then the object moves in a plane and the radius vector from the origin to the object sweeps out area at a constant rate. This is the equal area rule. In the context of planetary motion it is called Kepler's second law.

Lemma 17.3.1 says the object moves in a plane. From the assumption that the force field is a central force field, it follows from 12.5 that

$$2r'(t)\theta'(t) + r(t)\theta''(t) = 0$$

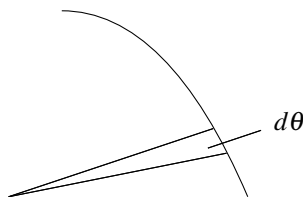
Multiply both sides of this equation by r . This yields

$$2rr'\theta' + r^2\theta'' = (r^2\theta')' = 0. \quad (17.1)$$

Consequently,

$$r^2\theta' = c \quad (17.2)$$

for some constant C . Now consider the following picture.



In this picture, $d\theta$ is the indicated angle and the two lines determining this angle are position vectors for the object at point t and point $t + dt$. The area of the sector, dA , is essentially $r^2 d\theta$ and so $dA = \frac{1}{2} r^2 d\theta$. Therefore,

$$\frac{dA}{dt} = \frac{1}{2} r^2 \frac{d\theta}{dt} = \frac{c}{2}. \quad (17.3)$$

17.3.2 Inverse Square Law, Kepler's First Law

Consider the first of Kepler's laws, the one which states that planets move along ellipses. From Lemma 17.3.1, the motion is in a plane. Now from 12.5 and Newton's second law,

$$\begin{aligned} & (r''(t) - r(t)\theta'(t)^2) \mathbf{e}_r + (2r'(t)\theta'(t) + r(t)\theta''(t)) \mathbf{e}_\theta \\ &= -\frac{GMm}{m} \left(\frac{1}{r^2} \right) \mathbf{e}_r = -k \left(\frac{1}{r^2} \right) \mathbf{e}_r \end{aligned}$$

Thus $k = GM$ and

$$r''(t) - r(t)\theta'(t)^2 = -k \left(\frac{1}{r^2} \right), \quad 2r'(t)\theta'(t) + r(t)\theta''(t) = 0. \quad (17.4)$$

As in 17.1, $(r^2\theta')' = 0$ and so there exists a constant c , such that

$$r^2\theta' = c. \quad (17.5)$$

Now the other part of 17.4 and 17.5 implies

$$r''(t) - r(t)\theta'(t)^2 = r''(t) - r(t) \left(\frac{c^2}{r^4} \right) = -k \left(\frac{1}{r^2} \right). \quad (17.6)$$

It is only r as a function of θ which is of interest. Using the chain rule,

$$r' = \frac{dr}{d\theta} \frac{d\theta}{dt} = \frac{dr}{d\theta} \left(\frac{c}{r^2} \right) \quad (17.7)$$

and so also

$$\begin{aligned} r'' &= \frac{d^2r}{d\theta^2} \left(\frac{d\theta}{dt} \right) \left(\frac{c}{r^2} \right) + \frac{dr}{d\theta} (-2) (c) (r^{-3}) \frac{dr}{d\theta} \frac{d\theta}{dt} \\ &= \frac{d^2r}{d\theta^2} \left(\frac{c}{r^2} \right)^2 - 2 \left(\frac{dr}{d\theta} \right)^2 \left(\frac{c^2}{r^5} \right) \end{aligned} \quad (17.8)$$

Using 17.8 and 17.7 in 17.6 yields

$$\frac{d^2r}{d\theta^2} \left(\frac{c}{r^2} \right)^2 - 2 \left(\frac{dr}{d\theta} \right)^2 \left(\frac{c^2}{r^5} \right) - r(t) \left(\frac{c^2}{r^4} \right) = -k \left(\frac{1}{r^2} \right).$$

Now multiply both sides of this equation by r^4/c^2 to obtain

$$\frac{d^2r}{d\theta^2} - 2 \left(\frac{dr}{d\theta} \right)^2 \frac{1}{r} - r = \frac{-kr^2}{c^2}. \quad (17.9)$$

This is a nice differential equation for r as a function of θ but its solution is not clear. It turns out to be convenient to define a new dependent variable, $\rho \equiv r^{-1}$ so $r = \rho^{-1}$. Then

$$\frac{dr}{d\theta} = (-1)\rho^{-2} \frac{d\rho}{d\theta}, \quad \frac{d^2r}{d\theta^2} = 2\rho^{-3} \left(\frac{d\rho}{d\theta} \right)^2 + (-1)\rho^{-2} \frac{d^2\rho}{d\theta^2}.$$

Substituting this in to 17.9 yields

$$2\rho^{-3} \left(\frac{d\rho}{d\theta} \right)^2 + (-1)\rho^{-2} \frac{d^2\rho}{d\theta^2} - 2 \left(\rho^{-2} \frac{d\rho}{d\theta} \right)^2 \rho - \rho^{-1} = \frac{-k\rho^{-2}}{c^2}$$

which simplifies to

$$(-1)\rho^{-2} \frac{d^2\rho}{d\theta^2} - \rho^{-1} = \frac{-k\rho^{-2}}{c^2}$$

since those two terms which involve $\left(\frac{d\rho}{d\theta} \right)^2$ cancel. Now multiply both sides by $-\rho^2$ and this yields

$$\frac{d^2\rho}{d\theta^2} + \rho = \frac{k}{c^2}, \quad (17.10)$$

which is a much nicer differential equation. Let $R = \rho - \frac{k}{c^2}$. Then in terms of R , this differential equation is

$$\frac{d^2R}{d\theta^2} + R = 0.$$

Multiply both sides by $\frac{dR}{d\theta}$. Then using the chain rule,

$$\frac{1}{2} \frac{d}{d\theta} \left(\left(\frac{dR}{d\theta} \right)^2 + R^2 \right) = 0$$

and so

$$\left(\frac{dR}{d\theta}\right)^2 + R^2 = \delta^2 \quad (17.11)$$

for some $\delta > 0$. Therefore, there exists an angle $\psi = \psi(\theta)$ such that

$$R = \delta \sin(\psi), \quad \frac{dR}{d\theta} = \delta \cos(\psi)$$

because 17.11 says $(\frac{1}{\delta} \frac{dR}{d\theta}, \frac{1}{\delta} R)$ is a point on the unit circle. But differentiating, the first of the above equations,

$$\frac{dR}{d\theta} = \delta \cos(\psi) \frac{d\psi}{d\theta} = \delta \cos(\psi)$$

and so $\frac{d\psi}{d\theta} = 1$. Therefore, $\psi = \theta + \phi$. Choosing the coordinate system appropriately, you can assume $\phi = 0$. Therefore,

$$R = \rho - \frac{k}{c^2} = \frac{1}{r} - \frac{k}{c^2} = \delta \sin(\theta)$$

and so, solving for r ,

$$r = \frac{1}{\left(\frac{k}{c^2}\right) + \delta \sin \theta} = \frac{c^2/k}{1 + (c^2/k) \delta \sin \theta} = \frac{p\varepsilon}{1 + \varepsilon \sin \theta}$$

where

$$\varepsilon = (c^2/k) \delta \text{ and } p = c^2/k\varepsilon. \quad (17.12)$$

Here all these constants are nonnegative.

Thus

$$r + \varepsilon r \sin \theta = \varepsilon p$$

and so $r = (\varepsilon p - \varepsilon y)$. Then squaring both sides,

$$x^2 + y^2 = (\varepsilon p - \varepsilon y)^2 = \varepsilon^2 p^2 - 2p\varepsilon^2 y + \varepsilon^2 y^2$$

And so

$$x^2 + (1 - \varepsilon^2)y^2 = \varepsilon^2 p^2 - 2p\varepsilon^2 y. \quad (17.13)$$

In case $\varepsilon = 1$, this reduces to the equation of a parabola. If $\varepsilon < 1$, this reduces to the equation of an ellipse and if $\varepsilon > 1$, this is called a hyperbola. This proves that objects which are acted on only by a force of the form given in the above example move along hyperbolas, ellipses or circles. The case where $\varepsilon = 0$ corresponds to a circle. The constant ε is called the eccentricity. This is called Kepler's first law in the case of a planet.

17.3.3 Kepler's Third Law

Kepler's third law involves the time it takes for the planet to orbit the sun. From 17.13 you can complete the square and obtain

$$x^2 + (1 - \varepsilon^2) \left(y + \frac{p\varepsilon^2}{1 - \varepsilon^2} \right)^2 = \varepsilon^2 p^2 + \frac{p^2 \varepsilon^4}{(1 - \varepsilon^2)} = \frac{\varepsilon^2 p^2}{(1 - \varepsilon^2)},$$

and this yields

$$x^2 / \left(\frac{\varepsilon^2 p^2}{1 - \varepsilon^2} \right) + \left(y + \frac{p\varepsilon^2}{1 - \varepsilon^2} \right)^2 / \left(\frac{\varepsilon^2 p^2}{(1 - \varepsilon^2)^2} \right) = 1. \quad (17.14)$$

Now note this is the equation of an ellipse and that the diameter of this ellipse is

$$\frac{2\varepsilon p}{(1 - \varepsilon^2)} \equiv 2a. \quad (17.15)$$

This follows because

$$\frac{\varepsilon^2 p^2}{(1 - \varepsilon^2)^2} \geq \frac{\varepsilon^2 p^2}{1 - \varepsilon^2}.$$

Now let T denote the time it takes for the planet to make one revolution about the sun. It is left as an exercise for you to show that the area of an ellipse whose long axis is $2a$ and whose short axis is $2b$ is πab . This is an exercise in trig. substitutions and is a little tedious but routine. Using this formula, and 17.3 the following equation must hold.

$$\overbrace{\pi \frac{\varepsilon p}{\sqrt{1 - \varepsilon^2}} \frac{\varepsilon p}{(1 - \varepsilon^2)}}^{\text{area of ellipse}} = T \frac{c}{2}$$

Therefore,

$$T = \frac{2}{c} \frac{\pi \varepsilon^2 p^2}{(1 - \varepsilon^2)^{3/2}}$$

and so

$$T^2 = \frac{4\pi^2 \varepsilon^4 p^4}{c^2 (1 - \varepsilon^2)^3}$$

Now using 17.12, recalling that $k = GM$, and 17.15,

$$T^2 = \frac{4\pi^2 \varepsilon^4 p^4}{k \varepsilon p (1 - \varepsilon^2)^3} = \frac{4\pi^2 (\varepsilon p)^3}{k (1 - \varepsilon^2)^3} = \frac{4\pi^2 a^3}{k} = \frac{4\pi^2 a^3}{GM}.$$

Written more memorably, this has shown

$$T^2 = \frac{4\pi^2}{GM} \left(\frac{\text{diameter of ellipse}}{2} \right)^3. \quad (17.16)$$

This relationship is known as Kepler's third law.

17.4 The Angular Velocity Vector

Let $(\mathbf{i}(t), \mathbf{j}(t), \mathbf{k}(t))$ be a right handed system of unit basis vectors. Thus $\mathbf{k}(t) = \mathbf{i}(t) \times \mathbf{j}(t)$ and each vector has unit length. This represents a moving coordinate system. We assume that $\mathbf{i}(t), \mathbf{j}(t), \mathbf{k}(t)$ are each continuous having continuous derivatives, as many as needed for the following manipulations for t in some open interval. The various rules of differentiation of vector valued functions will be used to show the existence of an angular velocity vector.

Lemma 17.4.1 *The following hold. Whenever $\mathbf{r}(t), \mathbf{s}(t)$ are two vectors from the set of vectors $\{\mathbf{i}(t), \mathbf{j}(t), \mathbf{k}(t)\}$,*

$$\mathbf{r}(t) \cdot \mathbf{s}'(t) = -\mathbf{r}'(t) \cdot \mathbf{s}(t)$$

In particular, the case where $\mathbf{r} = \mathbf{s}$, implies $\mathbf{r}'(t) \cdot \mathbf{r}(t) = 0$.

Proof: By assumption, $\mathbf{r}(t) \cdot \mathbf{s}(t)$ is either 0 for all t or 1 in case $\mathbf{r} = \mathbf{s}$. Therefore, from the product rule,

$$\mathbf{r}(t) \cdot \mathbf{s}'(t) + \mathbf{r}'(t) \cdot \mathbf{s}(t) = 0$$

which yields the desired result. ■

Then the fundamental result is the following major theorem which gives the existence and uniqueness of the angular velocity vector.

Theorem 17.4.2 *Let $(\mathbf{i}(t), \mathbf{j}(t), \mathbf{k}(t))$ be a right handed orthogonal system of unit vectors as explained above. Then there exists a unique vector $\boldsymbol{\Omega}(t)$, the angular velocity vector, such that for $\mathbf{r}(t)$ any of the $\{\mathbf{i}(t), \mathbf{j}(t), \mathbf{k}(t)\}$,*

$$\mathbf{r}'(t) = \boldsymbol{\Omega}(t) \times \mathbf{r}(t)$$

Proof: First I will show that if this angular velocity vector $\boldsymbol{\Omega}(t)$ exists, then it must be of a certain form. This will prove uniqueness. After showing this, I will verify that it does what it needs to do by simply checking that it does so. In all considerations, recall that in the box product, the \times and \cdot can be switched. I will use this fact with no comment in what follows. So suppose that such an angular velocity vector exists. Then $\mathbf{i}'(t) = \boldsymbol{\Omega}(t) \times \mathbf{i}(t)$ with a similar formula holding for the other vectors. Also note that since this is a right handed system, $\mathbf{i}(t) \times \mathbf{j}(t) = \mathbf{k}(t)$, $\mathbf{j}(t) \times \mathbf{k}(t) = \mathbf{i}(t)$, and $\mathbf{k}(t) \times \mathbf{i}(t) = \mathbf{j}(t)$ as earlier. In addition, if you want the component of a vector \mathbf{v} with respect to some $\mathbf{r}(t)$, it is $\mathbf{v} \cdot \mathbf{r}(t) = v_r(t)$. Thus

$$\mathbf{v} = v_i \mathbf{i}(t) + v_j \mathbf{j}(t) + v_k \mathbf{k}(t), \quad v_r = \mathbf{v} \cdot \mathbf{r}(t) \text{ for each } \mathbf{r}(t) \in \{\mathbf{i}(t), \mathbf{j}(t), \mathbf{k}(t)\}$$

Then

$$\mathbf{i}'(t) \cdot \mathbf{j}(t) = \boldsymbol{\Omega}(t) \times \mathbf{i}(t) \cdot \mathbf{j}(t) = \boldsymbol{\Omega}(t) \cdot \mathbf{i}(t) \times \mathbf{j}(t) = \boldsymbol{\Omega}(t) \cdot \mathbf{k}(t) = \Omega_k(t)$$

Thus the component of $\boldsymbol{\Omega}(t)$ in the direction $\mathbf{k}(t)$ is determined. Next,

$$\mathbf{i}'(t) \cdot \mathbf{k}(t) = \boldsymbol{\Omega}(t) \times \mathbf{i}(t) \cdot \mathbf{k}(t) = \boldsymbol{\Omega}(t) \cdot \mathbf{i}(t) \times \mathbf{k}(t) = -\Omega_j(t)$$

and so the component in the direction $\mathbf{j}(t)$ is also determined. Next,

$$\mathbf{j}'(t) \cdot \mathbf{k}(t) = \boldsymbol{\Omega}(t) \times \mathbf{j}(t) \cdot \mathbf{k}(t) = \boldsymbol{\Omega}(t) \cdot (\mathbf{j}(t) \times \mathbf{k}(t)) = \Omega_i(t)$$

so the component of $\boldsymbol{\Omega}(t)$ in direction $\mathbf{i}(t)$ is determined. Thus, if there is such an angular velocity vector, it must be of the form

$$\boldsymbol{\Omega}(t) \equiv (\mathbf{j}'(t) \cdot \mathbf{k}(t)) \mathbf{i}(t) - (\mathbf{i}'(t) \cdot \mathbf{k}(t)) \mathbf{j}(t) + (\mathbf{i}'(t) \cdot \mathbf{j}(t)) \mathbf{k}(t)$$

It only remains to verify that this vector works. Recall Lemma 17.4.1 which will be used without comment in what follows. Does the above $\Omega(t)$ work?

$$\begin{aligned}\Omega(t) \times \mathbf{i}(t) &= (\mathbf{i}'(t) \cdot \mathbf{k}(t)) \mathbf{k}(t) \\ &\quad + (\mathbf{i}'(t) \cdot \mathbf{j}(t)) \mathbf{j}(t) + \left(\overbrace{\mathbf{i}'(t) \cdot \mathbf{i}(t)}^{=0} \right) \mathbf{i}(t) = \mathbf{i}'(t)\end{aligned}$$

$$\begin{aligned}\Omega(t) \times \mathbf{j}(t) &= (\mathbf{j}'(t) \cdot \mathbf{k}(t)) \mathbf{k}(t) + (\mathbf{i}'(t) \cdot \mathbf{j}(t)) (-\mathbf{i}(t)) \\ &= (\mathbf{j}'(t) \cdot \mathbf{k}(t)) \mathbf{k}(t) + (\mathbf{i}(t) \cdot \mathbf{j}'(t)) \mathbf{i}(t) = \mathbf{j}'(t)\end{aligned}$$

and finally,

$$\begin{aligned}\Omega(t) \times \mathbf{k}(t) &= (\mathbf{j}'(t) \cdot \mathbf{k}(t)) (-\mathbf{j}(t)) - (\mathbf{i}'(t) \cdot \mathbf{k}(t)) \mathbf{i}(t) \\ &= (\mathbf{j}(t) \cdot \mathbf{k}'(t)) \mathbf{j}(t) + (\mathbf{i}(t) \cdot \mathbf{k}'(t)) \mathbf{i}(t) = \mathbf{k}'(t)\end{aligned}$$

Thus, this $\Omega(t)$ is the angular velocity vector and there is only one. Of course it might have different descriptions but there can only be one and it is the vector just described. ■

This implies the following simple corollary.

Corollary 17.4.3 *Let $\mathbf{u}(t)$ be a vector such that its components with respect to the basis vectors $\mathbf{i}(t), \mathbf{j}(t), \mathbf{k}(t)$ are constant. Then $\mathbf{u}'(t) = \Omega(t) \times \mathbf{u}(t)$.*

Proof: Say $\mathbf{u}(t) = u_i \mathbf{i}(t) + u_j \mathbf{j}(t) + u_k \mathbf{k}(t)$. Then

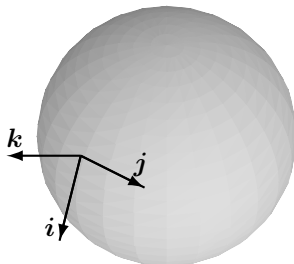
$$\begin{aligned}\mathbf{u}'(t) &= u_i \mathbf{i}'(t) + u_j \mathbf{j}'(t) + u_k \mathbf{k}'(t) \\ &= u_i \Omega(t) \times \mathbf{i}(t) + u_j \Omega(t) \times \mathbf{j}(t) + u_k \Omega(t) \times \mathbf{k}(t) \\ &= \Omega(t) \times (u_i \mathbf{i}(t) + u_j \mathbf{j}(t) + u_k \mathbf{k}(t)) = \Omega(t) \times \mathbf{u}(t) \quad \blacksquare\end{aligned}$$

17.5 Angular Velocity Vector on Earth

So how do you find the angular velocity vector? One way is to use the formula shown above. However, in important cases, this angular velocity vector can be determined from simple geometric reasoning. An obvious example concerns motion on the surface of the earth. Imagine you have a coordinate system fixed with the earth. Then it is actually rotating through space because the earth is turning. However, to an observer on the surface of the earth, these vectors are not moving and this observer wants to understand motion in terms of these apparently fixed vectors. This is a very interesting problem which can be understood relative to what was just discussed.

Of course the earth moves through space, but this is ignored because the accelerations relative to this motion are so small. After all, it takes a year to go around the sun.

Imagine a point on the surface of the earth which is not moving relative to the earth. Now consider unit vectors, one pointing South, one pointing East and one pointing directly away from the center of the earth.



Denote the first as $i(t)$, the second as $j(t)$, and the third as $k(t)$. If you are standing on the earth you will consider these vectors as fixed, but of course they are not. As the earth turns, they change direction and so each is in reality a function of t . What is the description of the angular velocity vector in this situation?

Let i^*, j^*, k^* be the usual basis vectors fixed in space with k^* pointing in the direction of the north pole from the center of the earth and let $i(t), j(t), k(t)$ be the unit vectors described earlier with $i(t)$ pointing South, $j(t)$ pointing East, and $k(t)$ pointing away from the center of the earth at some point of the rotating earth's surface $p(t)$. (This means that the components of $p(t)$ are constant with respect to the vectors fixed with the earth.) Letting $R(t)$ be the position vector of the point $p(t)$, from the center of the earth, observe that this is a typical vector having coordinates constant with respect to $i(t), j(t), k(t)$. Also, since the earth rotates from West to East and the speed of a point on the surface of the earth relative to an observer fixed in space is $\omega |R| \sin \phi$ where ω is the angular speed of the earth about an axis through the poles and ϕ is the polar angle measured from the positive z axis down as in spherical coordinates. It follows from the geometric definition of the cross product that

$$R'(t) = \omega k^* \times R(t)$$

Therefore, the vector of Theorem 17.4.2 is $\Omega(t) = \omega k^*$ because it acts like it should for vectors having components constant with respect to the vectors fixed with the earth. As mentioned, you could let θ, ρ, ϕ each be a function of t and use the formula above along with the chain rule to verify analytically that the angular velocity vector is what is claimed above. That is, you would have $\theta(t) = \omega t$ and the other spherical coordinates constant. See Problem 12 on Page 385 below for a more analytical explanation.

17.6 Coriolis Force and Centripetal Force

Let $p(t)$ be a point which has constant components relative to the moving coordinate system described above $\{i(t), j(t), k(t)\}$. For example, it could be a single point on the rotating earth or more generally simply a generic moving coordinate system. Let i^*, j^*, k^* be a typical rectangular coordinate system fixed in space and let $R(t)$ be the position vector of $p(t)$ from the origin fixed in space. In the case of the earth, think of the origin as the center of the earth. Thus the components of $R(t)$ with respect to the moving coordinate system are constants. A general observation is this. If $w(t) = w_1(t)i(t) + w_2(t)j(t) + w_3(t)k(t)$,

let $\mathbf{w}'_B(t)$ be

$$\mathbf{w}'_B(t) = w'_1(t) \mathbf{i}(t) + w'_2(t) \mathbf{j}(t) + w'_3(t) \mathbf{k}(t)$$

A dot will indicate the total derivative. Thus

$$\begin{aligned} \dot{\mathbf{w}}(t) &\equiv w'_1(t) \mathbf{i}(t) + w'_2(t) \mathbf{j}(t) + w'_3(t) \mathbf{k}(t) \\ &\quad + w_1(t) \dot{\mathbf{i}}(t) + w_2(t) \dot{\mathbf{j}}(t) + w_3(t) \dot{\mathbf{k}}(t) \\ &\equiv \mathbf{w}'_B(t) + \boldsymbol{\Omega}(t) \times \mathbf{w}(t) \end{aligned}$$

Thus, when you differentiate a vector \mathbf{w} , you write it as $\dot{\mathbf{w}} = \mathbf{w}'_B + \boldsymbol{\Omega}(t) \times \mathbf{w}$ where \mathbf{w}'_B is the perceived time derivative in the moving coordinate system.

$$\mathbf{w}'_B(t) = w'_1(t) \mathbf{i}(t) + w'_2(t) \mathbf{j}(t) + w'_3(t) \mathbf{k}(t)$$

Let $\mathbf{r}_B(t)$ be the position vector from this point $\mathbf{p}(t)$ to some other point.

$$\mathbf{r}_B(t) \equiv x(t) \mathbf{i}(t) + y(t) \mathbf{j}(t) + z(t) \mathbf{k}(t)$$

The acceleration perceived by an observer moving with the moving coordinate system would then be

$$\mathbf{a}_B(t) = x''(t) \mathbf{i}(t) + y''(t) \mathbf{j}(t) + z''(t) \mathbf{k}(t)$$

and the perceived velocity would be

$$\mathbf{v}_B(t) \equiv x'(t) \mathbf{i}(t) + y'(t) \mathbf{j}(t) + z'(t) \mathbf{k}(t)$$

Let $\mathbf{r}(t) \equiv \mathbf{R}(t) + \mathbf{r}_B(t)$. Then, since $\mathbf{R}(t)$ has constant components relative to the moving coordinate system, $\dot{\mathbf{R}}(t) = \boldsymbol{\Omega}(t) \times \mathbf{R}(t)$. It doesn't have constant components fixed in space, just with respect to the moving coordinate system. Thus, using the above observation with the usual conventions that \mathbf{v} is velocity and \mathbf{a} acceleration,

$$\mathbf{v}(t) = \dot{\mathbf{r}}(t) = \mathbf{r}'_B(t) + \boldsymbol{\Omega}(t) \times \mathbf{r}(t) \equiv \mathbf{v}_B(t) + \boldsymbol{\Omega}(t) \times \mathbf{r}(t)$$

Now go to the acceleration. Using the same process, $\mathbf{a} = \dot{\mathbf{v}} =$

$$\begin{aligned} \dot{\mathbf{v}}_B + \dot{\boldsymbol{\Omega}} \times \mathbf{r} + \boldsymbol{\Omega} \times \dot{\mathbf{r}} &= \mathbf{a}_B + \boldsymbol{\Omega} \times \mathbf{v}_B + \dot{\boldsymbol{\Omega}} \times \mathbf{r} + \boldsymbol{\Omega} \times (\mathbf{v}_B + \boldsymbol{\Omega} \times \mathbf{r}) \\ &= \mathbf{a}_B + 2\boldsymbol{\Omega} \times \mathbf{v}_B + \dot{\boldsymbol{\Omega}} \times \mathbf{r} + \boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{r}) \end{aligned} \quad (17.17)$$

17.7 Coriolis Force on the Rotating Earth

As explained above, on the rotating earth, $\boldsymbol{\Omega}$ is a constant and so 17.17 reduces to

$$\mathbf{a} = \mathbf{a}_B + 2(\boldsymbol{\Omega} \times \mathbf{v}_B) + \boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{r}) \quad (17.18)$$

Since $\mathbf{r}_B + \mathbf{R} = \mathbf{r}$,

$$\mathbf{a}_B = \mathbf{a} - \boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{R}) - 2\boldsymbol{\Omega} \times \mathbf{v}_B - \boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{r}_B). \quad (17.19)$$

In this formula, you can totally ignore the term $\boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{r}_B)$ because it is so small whenever you are considering motion near some point on the earth's surface. To see this, note

$\overbrace{\omega(24)(3600)}^{\text{seconds in a day}} = 2\pi$, and so $\omega = 7.2722 \times 10^{-5}$ in radians per second. If you are using seconds to measure time and feet to measure distance, this term is therefore, no larger than

$$\left(7.2722 \times 10^{-5}\right)^2 |r_B|.$$

Clearly this is not worth considering in the presence of the acceleration due to gravity which is approximately 32 feet per second squared near the surface of the earth.

If the acceleration \mathbf{a} is due to gravity, then

$$\begin{aligned} \mathbf{a}_B &= \mathbf{a} - \boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{R}) - 2\boldsymbol{\Omega} \times \mathbf{v}_B = \\ &= \overbrace{\frac{GM(\mathbf{R} + \mathbf{r}_B)}{|\mathbf{R} + \mathbf{r}_B|^3}}^{\equiv g} - \boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{R}) - 2\boldsymbol{\Omega} \times \mathbf{v}_B \equiv \mathbf{g} - 2\boldsymbol{\Omega} \times \mathbf{v}_B. \end{aligned}$$

Note that

$$\boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{R}) = (\boldsymbol{\Omega} \cdot \mathbf{R}) \boldsymbol{\Omega} - |\boldsymbol{\Omega}|^2 \mathbf{R}$$

and so \mathbf{g} , the acceleration relative to the moving coordinate system on the earth is not directed exactly toward the center of the earth except at the equator or at poles, although the components of acceleration which are in other directions are very small when compared with the acceleration due to the force of gravity and are often neglected. Therefore, if the only force acting on an object is due to gravity, the following formula describes the acceleration relative to a coordinate system moving with the earth's surface.

$$\mathbf{a}_B = \mathbf{g} - 2(\boldsymbol{\Omega} \times \mathbf{v}_B)$$

While the vector $\boldsymbol{\Omega}$ is quite small, if the relative velocity \mathbf{v}_B is large, the Coriolis acceleration could be significant. This is described in terms of the vectors $\mathbf{i}(t), \mathbf{j}(t), \mathbf{k}(t)$ next.

Letting (ρ, θ, ϕ) be the usual spherical coordinates of the point $\mathbf{p}(t)$ on the surface taken with respect to $\mathbf{i}^*, \mathbf{j}^*, \mathbf{k}^*$ the usual way with ϕ the polar angle, it follows the $\mathbf{i}^*, \mathbf{j}^*, \mathbf{k}^*$ coordinates of this point are

$$\begin{pmatrix} \rho \sin(\phi) \cos(\theta) \\ \rho \sin(\phi) \sin(\theta) \\ \rho \cos(\phi) \end{pmatrix}.$$

It follows,

$$\mathbf{i} = \cos(\phi) \cos(\theta) \mathbf{i}^* + \cos(\phi) \sin(\theta) \mathbf{j}^* - \sin(\phi) \mathbf{k}^*$$

$$\mathbf{j} = -\sin(\theta) \mathbf{i}^* + \cos(\theta) \mathbf{j}^* + 0 \mathbf{k}^*$$

and

$$\mathbf{k} = \sin(\phi) \cos(\theta) \mathbf{i}^* + \sin(\phi) \sin(\theta) \mathbf{j}^* + \cos(\phi) \mathbf{k}^*.$$

It is necessary to obtain \mathbf{k}^* in terms of the vectors, $\mathbf{i}(t), \mathbf{j}(t), \mathbf{k}(t)$ because, as shown earlier, $\omega \mathbf{k}^*$ is the angular velocity vector $\boldsymbol{\Omega}$. To simplify notation, I will suppress the dependence of these vectors on t . Thus the following equation needs to be solved for a, b, c to find $\mathbf{k}^* = a\mathbf{i} + b\mathbf{j} + c\mathbf{k}$

$$\overbrace{\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}}^{\mathbf{k}^*} = a \overbrace{\begin{pmatrix} \cos(\phi) \cos(\theta) \\ \cos(\phi) \sin(\theta) \\ -\sin(\phi) \end{pmatrix}}^{\mathbf{i}} + b \overbrace{\begin{pmatrix} -\sin(\theta) \\ \cos(\theta) \\ 0 \end{pmatrix}}^{\mathbf{j}} + c \overbrace{\begin{pmatrix} \sin(\phi) \cos(\theta) \\ \sin(\phi) \sin(\theta) \\ \cos(\phi) \end{pmatrix}}^{\mathbf{k}} \quad (17.20)$$

The solution is $a = -\sin(\phi)$, $b = 0$, and $c = \cos(\phi)$.

Now the Coriolis acceleration on the earth equals

$$2(\boldsymbol{\Omega} \times \mathbf{v}_B) = 2\omega \left(\overbrace{-\sin(\phi) \mathbf{i} + 0 \mathbf{j} + \cos(\phi) \mathbf{k}}^{\mathbf{k}^*} \right) \times (x' \mathbf{i} + y' \mathbf{j} + z' \mathbf{k}).$$

Recall that $\mathbf{i}, \mathbf{j}, \mathbf{k}$ is a right handed orthonormal system and so the method for finding the cross product is valid for these vectors. Thus, this equals

$$2\omega [(-y' \cos \phi) \mathbf{i} + (x' \cos \phi + z' \sin \phi) \mathbf{j} - (y' \sin \phi) \mathbf{k}]. \quad (17.21)$$

Remember ϕ is fixed and pertains to the fixed point $\mathbf{p}(t)$ on the earth's surface. Therefore, if the acceleration \mathbf{a} is due to gravity,

$$\mathbf{a}_B = \mathbf{g} - 2\omega [(-y' \cos \phi) \mathbf{i} + (x' \cos \phi + z' \sin \phi) \mathbf{j} - (y' \sin \phi) \mathbf{k}]$$

where $\mathbf{g} = -\frac{GM(\mathbf{R} + \mathbf{r}_B)}{|\mathbf{R} + \mathbf{r}_B|^3} - \boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{R})$ as explained above. The term $\boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{R})$ is quite small and will be neglected. However, the Coriolis force will not be neglected.

Example 17.7.1 Suppose a rock is dropped from a tall building. Where will it strike?

Assume $\mathbf{a} = -g\mathbf{k}$ and the \mathbf{j} component of \mathbf{a}_B is approximately

$$-2\omega (x' \cos \phi + z' \sin \phi).$$

The dominant term in this expression is clearly the second one because x' will be small. Also, the \mathbf{i} and \mathbf{k} contributions will be very small. Therefore, the following equation is descriptive of the situation.

$$\mathbf{a}_B = -g \mathbf{k} - 2z' \omega \sin \phi \mathbf{j}.$$

$z' = -gt$ approximately. Therefore, considering the \mathbf{j} component, this is

$$2gt\omega \sin \phi.$$

Two integrations give $(\omega g t^3 / 3) \sin \phi$ for the \mathbf{j} component of the relative displacement at time t .

This shows the rock does not fall directly towards the center of the earth as expected but slightly to the east.

17.8 The Foucault Pendulum*

In 1851 Foucault set a pendulum vibrating and observed the earth rotate out from under it. It was a very long pendulum with a heavy weight at the end so that it would vibrate for a long time without stopping². This is what allowed him to observe the earth rotate out from under it. Clearly such a pendulum will take 24 hours for the plane of vibration to appear to make one complete revolution at the north pole. It is also reasonable to expect that no such

²There is such a pendulum in the Eyring building at BYU and to keep people from touching it, there is a little sign which says Warning! 1000 ohms. You certainly don't want to encounter too many ohms! Most modern Foucault pendulums have a mechanism which applies a periodic force to keep it vibrating.

observed rotation would take place on the equator. Is it possible to predict what will take place at various latitudes?

Using 17.21, in 17.19,

$$\begin{aligned} \mathbf{a}_B &= \mathbf{a} - \boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{R}) \\ &\quad - 2\boldsymbol{\omega} [(-y' \cos \phi) \mathbf{i} + (x' \cos \phi + z' \sin \phi) \mathbf{j} - (y' \sin \phi) \mathbf{k}]. \end{aligned}$$

Neglecting the small term, $\boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{R})$, this becomes

$$= -g\mathbf{k} + \mathbf{T}/m - 2\boldsymbol{\omega} [(-y' \cos \phi) \mathbf{i} + (x' \cos \phi + z' \sin \phi) \mathbf{j} - (y' \sin \phi) \mathbf{k}]$$

where \mathbf{T} , the tension in the string of the pendulum, is directed towards the point at which the pendulum is supported, and m is the mass of the weight at the end of the pendulum. The pendulum can be thought of as the position vector from $(0, 0, l)$ to the surface of the sphere $x^2 + y^2 + (z - l)^2 = l^2$. Therefore,

$$\mathbf{T} = -T \frac{x}{l} \mathbf{i} - T \frac{y}{l} \mathbf{j} + T \frac{l - z}{l} \mathbf{k}$$

and consequently, the differential equations of relative motion are

$$\begin{aligned} x'' &= -T \frac{x}{ml} + 2\omega y' \cos \phi \\ y'' &= -T \frac{y}{ml} - 2\omega (x' \cos \phi + z' \sin \phi) \end{aligned}$$

and

$$z'' = T \frac{l - z}{ml} - g + 2\omega y' \sin \phi.$$

If the vibrations of the pendulum are small so that for practical purposes, $z'' = z = 0$, the last equation may be solved for T to get

$$gm - 2\omega y' \sin(\phi) m = T.$$

Therefore, the first two equations become

$$x'' = -(gm - 2\omega m y' \sin \phi) \frac{x}{ml} + 2\omega y' \cos \phi$$

and

$$y'' = -(gm - 2\omega m y' \sin \phi) \frac{y}{ml} - 2\omega (x' \cos \phi + z' \sin \phi).$$

All terms of the form xy' or $y'y$ can be neglected because it is assumed x and y remain small. Also, the pendulum is assumed to be long with a heavy weight so that x' and y' are also small. With these simplifying assumptions, the equations of motion become

$$x'' + g \frac{x}{l} = 2\omega y' \cos \phi$$

and

$$y'' + g \frac{y}{l} = -2\omega x' \cos \phi.$$

These equations are of the form

$$x'' + a^2 x = by', \quad y'' + a^2 y = -bx' \quad (17.22)$$

where $a^2 = \frac{g}{l}$ and $b = 2\omega \cos \phi$. There are systematic ways to solve the above linear system of ordinary differential equations, but for the purposes here, it is fairly tedious but routine to verify that for each constant c ,

$$x = c \sin\left(\frac{bt}{2}\right) \sin\left(\frac{\sqrt{b^2 + 4a^2}}{2}t\right), \quad y = c \cos\left(\frac{bt}{2}\right) \sin\left(\frac{\sqrt{b^2 + 4a^2}}{2}t\right) \quad (17.23)$$

yields a solution to 17.22 along with the initial conditions,

$$x(0) = 0, y(0) = 0, x'(0) = 0, y'(0) = \frac{c\sqrt{b^2 + 4a^2}}{2}. \quad (17.24)$$

It is clear from experiments with the pendulum that the earth does indeed rotate out from under it causing the plane of vibration of the pendulum to appear to rotate. The purpose of this discussion is not to establish this obvious fact but to predict how long it takes for the plane of vibration to make one revolution. There will be some instant in time at which the pendulum will be vibrating in a plane determined by \mathbf{k} and \mathbf{j} . (Recall \mathbf{k} points away from the center of the earth and \mathbf{j} points East.) At this instant in time, defined as $t = 0$, the conditions of 17.24 will hold for some value of c and so the solution to 17.22 having these initial conditions will be those of 17.23. (Some interesting mathematical details are being ignored here. Such initial value problems as 17.23 and 17.24 have only one solution so if you have found one, then you have found **the** solution. This is a general fact shown in differential equations courses. However, for the above system of equations see Problem 13 on Page 385 found below.) Writing these solutions differently,

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = c \begin{pmatrix} \sin\left(\frac{bt}{2}\right) \\ \cos\left(\frac{bt}{2}\right) \end{pmatrix} \sin\left(\frac{\sqrt{b^2 + 4a^2}}{2}t\right)$$

This is very interesting! The vector, $c \begin{pmatrix} \sin\left(\frac{bt}{2}\right) \\ \cos\left(\frac{bt}{2}\right) \end{pmatrix}$ always has magnitude equal to $|c|$ but its direction changes very slowly because b is very small. The plane of vibration is determined by this vector and the vector \mathbf{k} . The term $\sin\left(\frac{\sqrt{b^2 + 4a^2}}{2}t\right)$ changes relatively fast and takes values between -1 and 1 . This is what describes the actual observed vibrations of the pendulum. Thus the plane of vibration will have made one complete revolution when $t = T$ for $\frac{bT}{2} \equiv 2\pi$. Therefore, the time it takes for the earth to turn out from under the pendulum is

$$T = \frac{4\pi}{2\omega \cos \phi} = \frac{2\pi}{\omega} \sec \phi.$$

Since ω is the angular speed of the rotating earth, it follows $\omega = \frac{2\pi}{24} = \frac{\pi}{12}$ in radians per hour. Therefore, the above formula implies $T = 24 \sec \phi$. I think this is really amazing. You could determine latitude, not by taking readings with instruments using the North star but by doing an experiment with a big pendulum. You would set it vibrating, observe T in hours, and then solve the above equation for ϕ . Also note the pendulum would not appear to change its plane of vibration at the equator because $\lim_{\phi \rightarrow \pi/2} \sec \phi = \infty$.

17.9 Exercises

1. Find the length of the cardioid, $r = 1 + \cos \theta$, $\theta \in [0, 2\pi]$. **Hint:** A parametrization is $x(\theta) = (1 + \cos \theta) \cos \theta$, $y(\theta) = (1 + \cos \theta) \sin \theta$.

2. In general, show that the length of the curve given in polar coordinates by $r = f(\theta)$, $\theta \in [a, b]$ equals $\int_a^b \sqrt{f'(\theta)^2 + f(\theta)^2} d\theta$.
3. Using the above problem, find the lengths of graphs of the following polar curves.
 - (a) $r = \theta$, $\theta \in [0, 3]$
 - (b) $r = 2 \cos \theta$, $\theta \in [-\pi/2, \pi/2]$
 - (c) $r = 1 + \sin \theta$, $\theta \in [0, \pi/4]$
 - (d) $r = e^\theta$, $\theta \in [0, 2]$
 - (e) $r = \theta + 1$, $\theta \in [0, 1]$

4. Suppose the curve given in polar coordinates by $r = f(\theta)$ for $\theta \in [a, b]$ is rotated about the y axis. Find a formula for the resulting surface of revolution. You should get

$$2\pi \int_a^b f(\theta) \cos(\theta) \sqrt{f'(\theta)^2 + f(\theta)^2} d\theta$$

5. Using the result of the above problem, find the area of the surfaces obtained by revolving the polar graphs about the y axis.
 - (a) $r = \theta \sec(\theta)$, $\theta \in [0, 2]$
 - (b) $r = 2 \cos \theta$, $\theta \in [-\pi/2, \pi/2]$
 - (c) $r = e^\theta$, $\theta \in [0, 2]$
 - (d) $r = (1 + \theta) \sec(\theta)$, $\theta \in [0, 1]$
6. Suppose an object moves in such a way that $r^2 \theta'$ is a constant. Show that the only force acting on the object is a central force.
7. Explain why low pressure areas rotate counter clockwise in the Northern hemisphere and clockwise in the Southern hemisphere. **Hint:** Note that from the point of view of an observer fixed in space above the North pole, the low pressure area already has a counter clockwise rotation because of the rotation of the earth and its spherical shape. Now consider 17.2. In the low pressure area stuff will move toward the center so r gets smaller. How are things different in the Southern hemisphere?
8. What are some physical assumptions which are made in the above derivation of Kepler's laws from Newton's laws of motion?
9. The orbit of the earth is pretty nearly circular and the distance from the sun to the earth is about 149×10^6 kilometers. Using 17.16 and the above value of the universal gravitation constant, determine the mass of the sun. The earth goes around it in 365 days. (Actually it is 365.256 days.)
10. It is desired to place a satellite above the equator of the earth which will rotate about the center of mass of the earth every 24 hours. Is it necessary that the orbit be circular? What if you want the satellite to stay above the same point on the earth at all times? If the orbit is to be circular and the satellite is to stay above the same point, at what distance from the center of mass of the earth should the satellite be? You may use that the mass of the earth is 5.98×10^{24} kilograms. Such a satellite is called geosynchronous.

11. Show directly that the area of the inside of an ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ is πab . **Hint:** Solve for y and consider the top half of the ellipse.
12. Recall the formula derived above for the angular velocity vector

$$\boldsymbol{\Omega}(t) = (\mathbf{j}'(t) \cdot \mathbf{k}(t)) \mathbf{i}(t) - (\mathbf{i}'(t) \cdot \mathbf{k}(t)) \mathbf{j}(t) + (\mathbf{i}'(t) \cdot \mathbf{j}(t)) \mathbf{k}(t)$$

In the case of the rotating earth, $\mathbf{i}(t)$, $\mathbf{j}(t)$, and $\mathbf{k}(t)$ are respectively

$$\begin{pmatrix} \cos(\omega t) \cos \phi \\ \cos \phi \sin(\omega t) \\ -\sin \phi \end{pmatrix}, \begin{pmatrix} -\sin(\omega t) \\ \cos(\omega t) \\ 0 \end{pmatrix}, \begin{pmatrix} \sin(\phi) \cos(\omega t) \\ \sin(\phi) \sin(\omega t) \\ \cos(\phi) \end{pmatrix}$$

where column vectors are in terms of the fixed vectors \mathbf{i}^* , \mathbf{j}^* , \mathbf{k}^* . Show directly that $\boldsymbol{\Omega}(t) = \omega \mathbf{k}^*$ as claimed above.

13. Suppose you have

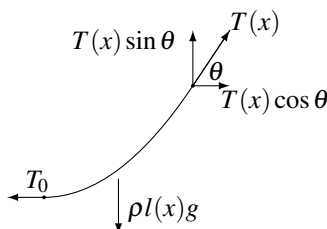
$$x'' + a^2 x = by', \quad y'' + a^2 y = -bx' \quad (17.25)$$

and $x(0) = x'(0) = y(0) = y'(0) = 0$. Show that $x(t) = y(t) = 0$. Show this implies there is only one solution to the initial value problem 17.23 and 17.24. **Hint:** If you had two solutions to 17.23 and 17.24, \tilde{x}, \tilde{y} and \hat{x}, \hat{y} , consider $x = \hat{x} - \tilde{x}$ and $y = \hat{y} - \tilde{y}$ and show x, y satisfies 17.25. To show the first part, multiply the first equation by x' the second by y' add and obtain the following using the product rule.

$$\frac{d}{dt} \left((x')^2 + (y')^2 + a^2 (x^2 + y^2) \right) = 0$$

Thus the inside is a constant. From the initial condition, this constant can only be 0.

14. This problem is about finding the equation of a hanging chain. Consider the following picture of a portion of this chain.



In this picture, ρ denotes the density of the chain which is assumed to be constant and g is the acceleration due to gravity. $T(x)$ and T_0 represent the magnitude of the tension in the chain at x and at 0 respectively, as shown and $l(x)$ is the length of the chain up to x . Let the bottom of the chain be at the origin as shown. If this chain does not move, then all these forces acting on it must balance. In particular,

$$T(x) \sin \theta = l(x) \rho g, \quad T(x) \cos \theta = T_0.$$

Therefore, dividing these yields $\tan(\theta) = \frac{\sin \theta}{\cos \theta} = l(x) \overbrace{\rho g / T_0}^{\equiv c}$. Now letting $y(x)$ denote the y coordinate of the hanging chain corresponding to x , $\tan \theta = y'(x)$. Therefore, this yields $y'(x) = cl(x)$. From formula for the length of a graph, explain why

$l'(x) = \sqrt{1 + y'(x)^2}$. Explain why $y''(x) = cl'(x) = c\sqrt{1 + y'(x)^2}$. Now let $z(x) = y'(x)$ and explain why $\frac{z'(x)}{\sqrt{1+z^2}} = c$. Therefore, $\int \frac{z'(x)}{\sqrt{1+z^2}} dx = cx + d$. Change variables and verify that $\sinh^{-1}(y'(x)) = cx + d$. Now verify that $y(x) = \frac{1}{c} \cosh(cx + d) + k$ which is the equation of a catenary.

Part II

Functions of Many Variables

Chapter 18

Linear Functions

Calculus of many variables involves the consideration of functions of many variables, just as calculus of one variable, considered earlier is about functions of one variable. Recall this could involve a function which has vector values, but the function depended on only one variable. When you consider functions of many variables, the easiest are those which are linear.

18.1 The Matrix of a Linear Transformation

The next definition is on what it means for a function to be linear.

Definition 18.1.1 Let $T : \mathbb{R}^n \rightarrow \mathbb{R}^m$. This function is **linear** if whenever α, β are numbers and \mathbf{x}, \mathbf{y} are vectors in \mathbb{R}^n , it follows that

$$T(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha T(\mathbf{x}) + \beta T(\mathbf{y})$$

Such linear functions are also called linear transformations or linear maps. Also, for $T : \mathbb{R}^n \rightarrow \mathbb{R}^m$ linear, it is standard to write $T \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$.

The first thing we need to do is to give an easily usable description of such a linear function. This will involve special vectors called \mathbf{e}_k . Also, from now on bold face \mathbf{x} will refer to a vector as earlier but now the vector will be written as a column vector. Thus

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

would denote a vector in \mathbb{R}^n . Actually, to save space, this vector will often be written as $(x_1 \ \cdots \ x_n)^T$ the exponent T indicating that one is to make this row of numbers into a column of numbers as above. All other conventions about adding and multiplying by numbers (scalars) are the same as discussed earlier. Now for the definition of the special vectors \mathbf{e}_k ,

Definition 18.1.2 \mathbf{e}_k is the vector $(0 \ \cdots \ 1 \ \cdots \ 0)^T$ where there is a 1 in the k^{th} position and a 0 in every other position. Thus if $\mathbf{x} = (x_1 \ \cdots \ x_n)^T$ is a vector

in \mathbb{R}^n ,

$$\mathbf{x} = x_1 \mathbf{e}_1 + x_2 \mathbf{e}_2 + \cdots + x_n \mathbf{e}_n$$

Written out, this is of the form

$$\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = x_1 \begin{pmatrix} 1 \\ \vdots \\ 0 \end{pmatrix} + \cdots + x_n \begin{pmatrix} 0 \\ \vdots \\ 1 \end{pmatrix}$$

As an example,

$$\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = 1 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + 2 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + 3 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

Be sure you understand this before reading further. You need to use the rules of vector addition and scalar multiplication discussed earlier but this time applied to column vectors.

Proposition 18.1.3 Let $T : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be linear, $T \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$. Then for

$$\mathbf{x} = \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix}^T,$$

$$T(\mathbf{x}) = x_1 T(\mathbf{e}_1) + x_2 T(\mathbf{e}_2) + \cdots + x_n T(\mathbf{e}_n)$$

In other words, for each $i \leq m$,

$$T(\mathbf{x})_i = x_1 T(\mathbf{e}_1)_i + x_2 T(\mathbf{e}_2)_i + \cdots + x_n T(\mathbf{e}_n)_i \equiv \sum_k x_k T(\mathbf{e}_k)_i$$

Proof: Since T is linear, $T(\mathbf{x}) = T(\sum_{k=1}^n x_k \mathbf{e}_k) = \sum_{k=1}^n x_k T(\mathbf{e}_k)$ which is the above. ■

Note that Proposition 18.1.3 shows that if you know what the linear function does to each \mathbf{e}_k , then you know what it does to an arbitrary vector \mathbf{x} .

Example 18.1.4 Suppose $T : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ is linear and

$$T\mathbf{e}_1 \equiv \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, T\mathbf{e}_2 \equiv \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, T\mathbf{e}_3 \equiv \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$$

Describe $T(\mathbf{x})$.

According to the above proposition,

$$T\mathbf{x} = x_1 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + x_2 \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} + x_3 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$$

There is a shortened version of this described in the following definition.

Definition 18.1.5 Letting \mathbf{x} be a vector in \mathbb{R}^n , and letting $\mathbf{u}_1, \dots, \mathbf{u}_n$ be vectors in \mathbb{R}^m , the **linear combination**

$$x_1 \mathbf{u}_1 + x_2 \mathbf{u}_2 + \cdots + x_n \mathbf{u}_n \equiv \sum_{k=1}^n x_k \mathbf{u}_k$$

is written as

$$\begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_n \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

Here $\begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_n \end{pmatrix}$ is called an $m \times n$ matrix, meaning it is a rectangular array of numbers having m rows (rows are horizontal) and n columns (columns are vertical). The k^{th} column from the left will be \mathbf{u}_k . Note that a linear combination is just an expression consisting of scalars times vectors added together. For T a linear transformation, its matrix A is such that $T(\mathbf{x}) = A\mathbf{x}$.

Theorem 18.1.6 Let $T \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$. Then the matrix of T is the following where each $T\mathbf{e}_k$ is a column vector:

$$\begin{pmatrix} T\mathbf{e}_1 & T\mathbf{e}_2 & \cdots & T\mathbf{e}_n \end{pmatrix}$$

Proof: This follows from the above definition and Proposition 18.1.3. ■

Example 18.1.7 Write the following as a matrix times a vector.

$$2 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - 2 \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} + 3 \begin{pmatrix} 0 \\ -4 \\ 2 \end{pmatrix}$$

According to the above definition, this is of the form

$$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & -4 \\ 1 & 3 & 2 \end{pmatrix} \begin{pmatrix} 2 \\ -2 \\ 3 \end{pmatrix}$$

When you multiply a matrix times a vector, you are just specifying a linear combination of the columns of the matrix. Thus every linear function $T : \mathbb{R}^n \rightarrow \mathbb{R}^m$ can be written as follows:

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

where A is an $m \times n$ matrix called the matrix of the linear transformation. This matrix is denoted as $[T]$. This is formalized in the following definition.

Definition 18.1.8 Let A be an $m \times n$ matrix. Then A_{ij} will denote the number in the i^{th} row and j^{th} column. $[T]$ denotes the $m \times n$ matrix such that $T(\mathbf{x}) = [T]\mathbf{x}$.

Example 18.1.9 Say $A = \begin{pmatrix} 1 & 2 & -5 \\ 4 & -7 & 2 \end{pmatrix}$. Then $A_{11} = 1, A_{12} = 2, A_{23} = 2, A_{22} = -7$ etc.

When writing A_{ij} the first index i always refers to the row and the second listed index refers to the column. This is hard for some of us to remember. Perhaps it will help to think **Row**man **Cath**olic. Another thing which is sometimes hard to remember is that the columns are vertical like those on the Parthenon in Athens and the rows are horizontal like the rows made by a tractor pulling a plow.

Definition 18.1.10 Suppose $T : \mathbb{R}^n \rightarrow \mathbb{R}^m$ and $S : \mathbb{R}^m \rightarrow \mathbb{R}^p$. You could consider the composition of these functions $S \circ T$ defined as $S \circ T(\mathbf{x}) \equiv S(T(\mathbf{x}))$.

With this definition, which is really nothing more than a re-statement of definitions from pre-calculus or algebra, the following is a fundamental theorem. It says that, appropriately defined, matrix multiplication corresponds to composition of linear transformations. This definition will be as follows.

Definition 18.1.11 Let A be an $m \times n$ matrix and let B be an $n \times p$ matrix. Then

$$(AB)_{ij} = \sum_{k=1}^n A_{ik}B_{kj} = A_{i1}B_{1j} + A_{i2}B_{2j} + \cdots + A_{in}B_{nj}$$

In terms of familiar concepts, the ij^{th} entry of AB is the i^{th} row of A times the j^{th} column of B meaning you take the dot product of the i^{th} row of A with the j^{th} column of B . Note that $A\mathbf{x}$ is a special case of this. Indeed,

$$(A\mathbf{x})_i = \sum_k A_{ik}x_k$$

This next theorem shows that this is what is needed in order to have matrix multiplication correspond to composition of linear transformations.

Theorem 18.1.12 Let $T : \mathbb{R}^n \rightarrow \mathbb{R}^m$ and $S : \mathbb{R}^m \rightarrow \mathbb{R}^p$ and suppose both T and S are linear. Then $S \circ T : \mathbb{R}^n \rightarrow \mathbb{R}^p$ is also linear and

$$[S \circ T] = [S][T]$$

where matrix multiplication is defined in Definition 18.1.11.

Proof: By definition,

$$\begin{aligned} \sum_j [S \circ T]_{ij} x_j &\equiv (S \circ T(\mathbf{x}))_i = (S(T(\mathbf{x})))_i = ([S](T\mathbf{x}))_i \\ &= \sum_k [S]_{ik} (T\mathbf{x})_k = \sum_k [S]_{ik} ([T]\mathbf{x})_k \\ &= \sum_k [S]_{ik} \sum_j [T]_{kj} x_j = \sum_j \left(\sum_k [S]_{ik} [T]_{kj} \right) x_j \end{aligned}$$

It follows, since \mathbf{x} is completely arbitrary that for each i , and for each j ,

$$[S \circ T]_{ij} = \sum_k [S]_{ik} [T]_{kj} \blacksquare$$

Here is something you must understand about matrix multiplication. For A and B matrices, in order to form the product AB the number of columns of A must equal the number of rows of B .

$$(m \times n)(n \times p) = m \times p, (m \times n)(k \times p) = \text{nonsense} \quad (18.1)$$

The two outside numbers give the size of the product and the middle two numbers must match. You must have the same number of columns on the left as you have rows on the right.

Example 18.1.13 Let $A = \begin{pmatrix} 1 & -1 & 2 \\ 3 & -2 & 1 \end{pmatrix}$ and $B = \begin{pmatrix} 2 & 3 \\ -1 & 1 \\ 0 & 3 \end{pmatrix}$. Then find AB . After this, find BA

Consider first AB . It is the product of a 2×3 and a 3×2 matrix and so it is a 2×2 matrix. The top left corner is the dot product of the top row of A and the first column of B and so forth. Be sure you can show the following that $AB = \begin{pmatrix} 3 & 8 \\ 8 & 10 \end{pmatrix}$, $BA = \begin{pmatrix} 11 & -8 & 7 \\ 2 & -1 & -1 \\ 9 & -6 & 3 \end{pmatrix}$.

Note this shows that matrix multiplication is not commutative. Indeed, it can result in matrices of different size when you interchange the order. Here is a perplexing little observation. If you add the entries on the main diagonal of both matrices in the above, you get the same number 13. This is the diagonal from upper left to lower right. You might wonder whether this always happens or if this is just a fluke. In fact, it will always happen. You should try and show this.

You can add matrices of the same size by adding the corresponding entries. Indeed, you must do this if you want to preserve the idea that matrix multiplication of a vector gives a linear transformation of the vector. Say $T, S \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$. These are just functions of a special sort. Thus $T + S$ is defined as the function which does the following: $(T + S)(\mathbf{x}) \equiv T(\mathbf{x}) + S(\mathbf{x})$.

$$\begin{aligned} \sum_k [T + S]_{ik} x_k &\equiv ((T + S)(\mathbf{x}))_i \equiv (T(\mathbf{x}) + S(\mathbf{x}))_i \\ &= (T(\mathbf{x}))_i + (S(\mathbf{x}))_i = ([T]\mathbf{x})_i + ([S]\mathbf{x})_i \\ &= \sum_k [T]_{ik} x_k + \sum_k [S]_{ik} x_k = \sum_k ([T]_{ik} + [S]_{ik}) x_k \end{aligned}$$

Since \mathbf{x} is arbitrary, it follows that $[T + S]_{ik} = [T]_{ik} + [S]_{ik}$. In other words, you must add corresponding entries. This shows why you must add matrices of the same size. Similarly you need $\alpha[T] = [\alpha T]$.

Then in terms scalar multiplication and addition of either matrices or linear transformations, following properties are called the vector space axioms.

- Commutative Law Of Addition.

$$A + B = B + A, \quad (18.2)$$

- Associative Law for Addition.

$$(A + B) + C = A + (B + C), \quad (18.3)$$

- Existence of an Additive Identity

$$A + 0 = A, \quad (18.4)$$

- Existence of an Additive Inverse

$$A + (-A) = 0, \quad (18.5)$$

Also for α, β scalars, the following additional properties hold.

- Distributive law over Matrix Addition.

$$\alpha(A+B) = \alpha A + \alpha B, \quad (18.6)$$

- Distributive law over Scalar Addition

$$(\alpha + \beta)A = \alpha A + \beta A, \quad (18.7)$$

- Associative law for Scalar Multiplication

$$\alpha(\beta A) = \alpha\beta(A), \quad (18.8)$$

- Rule for Multiplication by 1.

$$1A = A. \quad (18.9)$$

Example 18.1.14 $\begin{pmatrix} 1 & 2 & 3 \\ 4 & -5 & -8 \end{pmatrix} + \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} = \begin{pmatrix} 2 & 4 & 6 \\ 8 & 0 & -2 \end{pmatrix}.$

Example 18.1.15 Find $\begin{pmatrix} 1 & 2 & 3 \\ 4 & -5 & -8 \end{pmatrix} \begin{pmatrix} 2 & 8 \\ 1 & 0 \\ 2 & -2 \end{pmatrix} + \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$

$$\begin{aligned} & \begin{pmatrix} 1 & 2 & 3 \\ 4 & -5 & -8 \end{pmatrix} \begin{pmatrix} 2 & 8 \\ 1 & 0 \\ 2 & -2 \end{pmatrix} + \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 10 & 2 \\ -13 & 48 \end{pmatrix} + \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} 11 & 4 \\ -11 & 49 \end{pmatrix} \end{aligned}$$

Although matrix multiplication (composition of linear transformations) is not commutative, it does have several very important properties.

Proposition 18.1.16 *If all multiplications and additions make sense, the following hold for matrices A, B, C and a, b scalars.*

$$A(aB + bC) = a(AB) + b(AC) \quad (18.10)$$

$$(B + C)A = BA + CA \quad (18.11)$$

$$A(BC) = (AB)C \quad (18.12)$$

Proof: Using the definition for matrix multiplication, $(A(aB + bC))_{ij} =$

$$\begin{aligned} \sum_k A_{ik}(aB + bC)_{kj} &= \sum_k A_{ik}(aB_{kj} + bC_{kj}) = a \sum_k A_{ik}B_{kj} + b \sum_k A_{ik}C_{kj} \\ &= a(AB)_{ij} + b(AC)_{ij} = (a(AB) + b(AC))_{ij}. \end{aligned}$$

Thus $A(B + C) = AB + AC$ as claimed. Formula 18.11 is entirely similar.

Formula 18.12 is the associative law of multiplication. Using Definition 18.1.11,

$$\begin{aligned}(A(BC))_{ij} &= \sum_k A_{ik} (BC)_{kj} = \sum_k A_{ik} \sum_l B_{kl} C_{lj} \\ &= \sum_l (AB)_{il} C_{lj} = ((AB)C)_{ij}.\end{aligned}$$

This proves 18.12. ■

Specializing 18.10 to the case where B, C are vectors, this shows that $x \rightarrow Ax$ is a linear transformation. Thus every linear transformation can be realized by matrix multiplication and conversely, if you consider matrix multiplication, this is a linear transformation. This is why in this book, I will emphasize matrix multiplication rather than the abstract concept of a linear transformation.

Also note that 18.12, along with the theorem that matrix multiplication corresponds to composition of linear transformations, follows from the general observation from college algebra that

$$S \circ (T \circ V) = (S \circ T) \circ V$$

As to the restriction 18.1, it is essentially the statement that if you want $S \circ T$, then the possible values of T must be in the domain of S .

Definition 18.1.17 Let A be a $m \times n$ matrix. Then A^T is the $n \times m$ matrix defined as $(A^T)_{ij} \equiv A_{ji}$. In other words, the i^{th} row becomes the i^{th} column.

Example 18.1.18 Let $A = \begin{pmatrix} 1 & 4 & -6 \\ -3 & 2 & 1 \end{pmatrix}$. Then $A^T = \begin{pmatrix} 1 & -3 \\ 4 & 2 \\ -6 & 1 \end{pmatrix}$.

There is a fundamental theorem about how the transpose relates to multiplication.

Lemma 18.1.19 Let A be an $m \times n$ matrix and let B be a $n \times p$ matrix. Then

$$(AB)^T = B^T A^T \quad (18.13)$$

and if α and β are scalars,

$$(\alpha A + \beta B)^T = \alpha A^T + \beta B^T \quad (18.14)$$

Proof: From the definition,

$$\left((AB)^T \right)_{ij} = (AB)_{ji} = \sum_k A_{jk} B_{ki} = \sum_k (B^T)_{ik} (A^T)_{kj} = (B^T A^T)_{ij}$$

The proof of Formula 18.14 is left as an exercise and this proves the lemma. ■

Definition 18.1.20 An $n \times n$ matrix, A is said to be **symmetric** if $A = A^T$. It is said to be **skew symmetric** if $A = -A^T$.

Example 18.1.21 $\begin{pmatrix} 2 & 1 & 3 \\ 1 & 5 & -3 \\ 3 & -3 & 7 \end{pmatrix}$ is symmetric and $\begin{pmatrix} 0 & 1 & 3 \\ -1 & 0 & 2 \\ -3 & -2 & 0 \end{pmatrix}$ is skew symmetric.

Example 18.1.22 Find $A^T B + C^T$ where $A = \begin{pmatrix} 1 & 2 \end{pmatrix}$, $B = \begin{pmatrix} 1 & 1 \end{pmatrix}$, $C = \begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix}$

$$\begin{pmatrix} 1 & 2 \end{pmatrix}^T \begin{pmatrix} 1 & 1 \end{pmatrix} + \begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix}^T = \begin{pmatrix} 1 & 1 \\ 2 & 2 \end{pmatrix} + \begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix}^T = \begin{pmatrix} 2 & 2 \\ 4 & 3 \end{pmatrix}$$

Example 18.1.23 For F an $m \times n$ matrix, it is *always* possible to do the multiplication $F^T F$.

This is true because F^T is $n \times m$ and F is $m \times n$.

Another important observation is the following which will be used frequently.

Proposition 18.1.24 Let A be an $m \times n$ matrix. Let $B = \begin{pmatrix} \mathbf{b}_1 & \cdots & \mathbf{b}_p \end{pmatrix}$ where each \mathbf{b}_k is a column vector or $n \times 1$ matrix. Then AB is an $m \times p$ matrix and

$$AB = \begin{pmatrix} A\mathbf{b}_1 & \cdots & A\mathbf{b}_p \end{pmatrix}$$

so the k^{th} column of AB is just $A\mathbf{b}_k$.

Proof: From the definition of multiplication of matrices, $(AB)_{ik} = \sum_r A_{ir} B_{rk}$. However,

$$\mathbf{b}_k = \begin{pmatrix} B_{1k} \\ \vdots \\ B_{nk} \end{pmatrix}$$

and so, from the way we multiply a matrix times a vector,

$$(A\mathbf{b}_k)_i = \sum_r A_{ir} (\mathbf{b}_k)_r = \sum_r A_{ir} B_{rk}$$

Thus, the i^{th} entry from the top of $A\mathbf{b}_k$ is the i^{th} entry in the k^{th} column of AB showing that indeed the claim is true. ■

18.2 Row Operations and Linear Equations

In Junior High, you learned to solve things like $ax = b$ when $a \neq 0$. The fundamental problem considered in this section is the higher dimensional version of this $A\mathbf{x} = \mathbf{b}$ where A is an $m \times n$ matrix. First of all, there might not even be a solution to this. Consider

$$\begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \end{pmatrix}$$

Obviously there is no solution because on the left you get $\begin{pmatrix} x+2y \\ x+2y \end{pmatrix}$ and you can't have $x+2y$ equal to both 1 and 3. In contrast to religion and liberal arts, we do not tolerate contradictory assertions in mathematics. When two or more equations result in such contradictions, we say the equations are inconsistent. When something like this happens, we say the solution is \emptyset the empty set. So how do you go about solving such equations when they can be solved or determining that there is no solution like the above? This involves the concept of a row operation.

Definition 18.2.1 *The row operations applied to a matrix A consist of the following*

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

It is very useful to show that each of these row operations can be accomplished by multiplication on the left by a suitable matrix called an elementary matrix. First is a definition of the identity matrix.

Definition 18.2.2 *An $n \times n$ matrix I is called the identity matrix if $I_{ij} = 1$ if $i = j$ and $I_{ij} = 0$ if $i \neq j$.*

The importance of the identity matrix is that when you multiply by it, nothing changes. It acts like 1.

Proposition 18.2.3 *Let A be an $m \times n$ matrix then if I is the $m \times m$ identity matrix, it follows that $IA = A$ and if I is the $n \times n$ identity matrix, then $AI = A$.*

Proof: From the definition of how we multiply matrices, $(IA)_{ij} = \sum_k I_{ik}A_{kj}$. Now each $I_{ik} = 0$ except when $k = i$ when it is 1. Hence the sum reduced so A_{ij} and so the ij^{th} entry of IA is the same as the ij^{th} entry of A and so $IA = A$ because they are the same matrix. On the other side it is similar and this is left as an exercise. ■

The identity matrix has 1 down the main diagonal and 0 everywhere else. This means it looks like this in the case of the 3×3 identity:

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

It is also standard notation to denote the ij^{th} entry of the identity matrix with the symbol δ_{ij} sometimes δ_j^i .

When you multiply by the identity, nothing happens, but when you multiply by an elementary matrix you end up doing a row operation. The next definition is what is meant by an elementary matrix.

Definition 18.2.4 *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 matrices¹.*

The importance of elementary matrices is that when you multiply on the left by one, it does the row operation which was used to produce the elementary matrix.

¹More generally, a permutation matrix is a matrix which comes by permuting the rows of the identity matrix, which means possibly more than two rows are switched.

Now consider what these elementary matrices look like. First consider the one which involves switching row i and row j where $i < j$. This matrix is of the form

$$\begin{pmatrix} \ddots & & & & \\ & 0 & & 1 & \\ & & \ddots & & \\ & 1 & & 0 & \\ & & & & \ddots \end{pmatrix}$$

Note how the i^{th} and j^{th} rows are switched in the identity matrix and there are thus all ones on the main diagonal except for those two positions indicated. The two exceptional rows are shown. The i^{th} row was the j^{th} and the j^{th} row was the i^{th} in the identity matrix. Now consider what this does to a column vector.

$$\begin{pmatrix} \ddots & & & & \\ & 0 & & 1 & \\ & & \ddots & & \\ & 1 & & 0 & \\ & & & & \ddots \end{pmatrix} \begin{pmatrix} \vdots \\ x_i \\ \vdots \\ x_j \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ x_j \\ \vdots \\ x_i \\ \vdots \end{pmatrix}$$

Now denote by P^{ij} the elementary matrix which comes from the identity from switching rows i and j . From what was just explained and Proposition 18.1.24,

$$P^{ij} \begin{pmatrix} \vdots & \vdots & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ip} \\ \vdots & \vdots & \vdots \\ a_{j1} & a_{j2} & \cdots & a_{jp} \\ \vdots & \vdots & \vdots \end{pmatrix} = \begin{pmatrix} \vdots & \vdots & \vdots \\ a_{j1} & a_{j2} & \cdots & a_{jp} \\ \vdots & \vdots & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ip} \\ \vdots & \vdots & \vdots \end{pmatrix}$$

This has established the following lemma.

Lemma 18.2.5 *Let P^{ij} denote the elementary matrix which involves switching the i^{th} and the j^{th} rows. Then*

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

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

Example 18.2.6 *Consider the following.*

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a & b \\ g & d \\ e & f \end{pmatrix} = \begin{pmatrix} g & d \\ a & b \\ e & f \end{pmatrix}$$

Next consider the row operation which involves multiplying the i^{th} row by a nonzero constant, c . The elementary matrix which results from applying this operation to the i^{th} row

of the identity matrix is of the form

$$\begin{pmatrix} \ddots & & & & 0 \\ & 1 & & & \\ & & c & & \\ & & & 1 & \\ 0 & & & & \ddots \end{pmatrix}$$

Now consider what this does to a column vector.

$$\begin{pmatrix} \ddots & & & & 0 \\ & 1 & & & \\ & & c & & \\ & & & 1 & \\ 0 & & & & \ddots \end{pmatrix} \begin{pmatrix} \vdots \\ v_{i-1} \\ v_i \\ v_{i+1} \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ v_{i-1} \\ cv_i \\ v_{i+1} \\ \vdots \end{pmatrix}$$

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 Proposition 18.1.24,

$$E(c, i) \begin{pmatrix} \vdots & \vdots & \vdots \\ a_{(i-1)1} & a_{(i-1)2} & \cdots & a_{(i-1)p} \\ a_{i1} & a_{i2} & \cdots & a_{ip} \\ a_{(i+1)1} & a_{(i+1)2} & \cdots & a_{(i+1)p} \\ \vdots & \vdots & \vdots \end{pmatrix} = \begin{pmatrix} \vdots & \vdots & \vdots \\ a_{(i-1)1} & a_{(i-1)2} & \cdots & a_{(i-1)p} \\ ca_{i1} & ca_{i2} & \cdots & ca_{ip} \\ a_{(i+1)1} & a_{(i+1)2} & \cdots & a_{(i+1)p} \\ \vdots & \vdots & \vdots \end{pmatrix}$$

This proves the following lemma.

Lemma 18.2.7 *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 .

Example 18.2.8 *Consider this.*

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \\ e & f \end{pmatrix} = \begin{pmatrix} a & b \\ 5c & 5d \\ e & f \end{pmatrix}$$

Finally consider the third of these row operations. Denote by $E(c \times i + j)$ the elementary matrix which replaces the j^{th} row with the j^{th} row added to c times the i^{th} row. In case $i < j$ this will be of the form

$$\begin{pmatrix} \ddots & & & & 0 \\ & 1 & & & \\ & & \ddots & & \\ & c & & 1 & \\ 0 & & & & \ddots \end{pmatrix}$$

Now consider what this does to a column vector.

$$\begin{pmatrix} \ddots & & & 0 \\ & 1 & & \\ & & \ddots & \\ c & & & 1 \\ 0 & & & & \ddots \end{pmatrix} \begin{pmatrix} \vdots \\ v_i \\ \vdots \\ v_j \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ v_i \\ \vdots \\ cv_i + v_j \\ \vdots \end{pmatrix}$$

Now from this and Proposition 18.1.24,

$$\begin{aligned} E(c \times i + j) \begin{pmatrix} \vdots & \vdots & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ip} \\ \vdots & \vdots & \vdots \\ a_{j1} & a_{j2} & \cdots & a_{jp} \\ \vdots & \vdots & \vdots \end{pmatrix} \\ = \begin{pmatrix} \vdots & \vdots & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ip} \\ \vdots & \vdots & \vdots \\ ca_{i1} + a_{j1} & ca_{i2} + a_{j2} & \cdots & ca_{ip} + a_{jp} \\ \vdots & \vdots & \vdots \end{pmatrix} \end{aligned}$$

The case where $i > j$ is handled similarly. This proves the following lemma.

Lemma 18.2.9 *Let $E(c \times i + j)$ denote the elementary matrix obtained from I by replacing the j^{th} row with c times the i^{th} row added to it. Then*

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

where B is obtained from A by replacing the j^{th} row of A with itself added to c times the i^{th} row of A .

Example 18.2.10 *Consider the third row operation.*

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & 0 & 1 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \\ e & f \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \\ 2a + e & 2b + f \end{pmatrix}$$

The next theorem is the main result.

Theorem 18.2.11 *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 . Furthermore, if E is an elementary matrix, then there is another elementary matrix \hat{E} such that $E\hat{E} = \hat{E}E = I$.*

Proof: The first part of this theorem has been proved in Lemmas 18.2.5 - 18.2.9. It only remains to verify the claim about the matrix \hat{E} . Consider first the elementary matrices corresponding to row operation of type three.

$$E(-c \times i + j)E(c \times i + j) = I.$$

This follows because the first matrix takes c times row i in the identity and adds it to row j . When multiplied on the left by $E(-c \times i + j)$ it follows from the first part of this theorem that you take the i^{th} row of $E(c \times i + j)$ which coincides with the i^{th} row of I since that row was not changed, multiply it by $-c$ and add to the j^{th} row of $E(c \times i + j)$ which was the j^{th} row of I added to c times the i^{th} row of I . Thus $E(-c \times i + j)$ multiplied on the left, undoes the row operation which resulted in $E(c \times i + j)$. The same argument applied to the product $E(c \times i + j)E(-c \times i + j)$ replacing c with $-c$ in the argument yields that this product is also equal to I . Therefore, there is an elementary matrix of the same sort which when multiplied by E on either side gives the identity.

Similar reasoning shows that for $E(c, i)$ the elementary matrix which comes from multiplying the i^{th} row by the nonzero constant c , you can take $\hat{E} = E((1/c), i)$.

Finally, consider P^{ij} which involves switching the i^{th} and the j^{th} rows $P^{ij}P^{ij} = I$ because by the first part of this theorem, multiplying on the left by P^{ij} switches the i^{th} and j^{th} rows of P^{ij} which was obtained from switching the i^{th} and j^{th} rows of the identity. First you switch them to get P^{ij} and then you multiply on the left by P^{ij} which switches these rows again and restores the identity matrix. ■

The way we solve the linear equation $Ax = b$ is to multiply on both sides by a succession of elementary matrices, in other words do row operations to both sides until the solution is obvious.

Proposition 18.2.12 *The solution set to $Ax = b$ is unchanged if the same row operation is done to A as to b . In other words, it has the same solution set as $EAx = Eb$.*

Proof: If x is such that $EAx = Eb$ then use the \hat{E} of the above Theorem 18.2.11 multiply both sides by \hat{E} and use the associative law to obtain $Ax = \hat{E}(EA)x = \hat{E}Eb = b$. If $Ax = b$, then $EAx = Eb$. Thus the two systems have the same solution set. ■

More generally, it is convenient to consider such a system in the form $(A|b)$ where the matrix A is on the left and there is another column b to give the last column. Such a matrix is called an augmented matrix. Then solving the system $Ax = b$ is equivalent to finding b as a linear combination of the columns of A . In other words, you want to find a linear relationship between b and the other columns. You are doing the same row operations on A as on b and so you might as well consider the system in this shortened form.

Example 18.2.13 *Solve*

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ -5 \end{pmatrix}$$

You consider

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & -1 & -5 \end{pmatrix}$$

Now proceed to do row operations. Take -1 times the top row and add to the bottom.

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & -2 & -6 \end{pmatrix}$$

Now take -1 times the bottom and add to the top.

$$\begin{pmatrix} 1 & 0 & 3 & 7 \\ 0 & 1 & -2 & -6 \end{pmatrix}$$

At this point it is obvious. Write as equations. You have $x + 3z = 7, y - 2z = -6$. You can therefore, pick z to be anything. I shall let it equal t . Then a solution is of the form

$$x = 7 - 3t, y = -6 + 2t, z = t, t \in \mathbb{R}$$

The solution is given parametrically in this form. Remember parametric lines. In this case, there is an infinite selection of solutions.

Example 18.2.14 Find the solution to

$$\begin{pmatrix} 2 & 2 & 3 \\ 1 & 1 & 0 \\ 2 & 2 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \\ 4 \end{pmatrix}$$

Do it the same.

$$\begin{pmatrix} 2 & 2 & 3 & 3 \\ 1 & 1 & 0 & 3 \\ 2 & 2 & 2 & 4 \end{pmatrix}$$

Now do row operations to this matrix to get

$$\begin{pmatrix} 1 & 1 & 0 & 3 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Thus $x = 3 - t, y = t, z = -1$ and $t \in \mathbb{R}$.

Example 18.2.15 Find the solution to $\begin{pmatrix} 1 & 2 & 1 \\ 0 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}$.

Add in the last column as above.

$$\begin{pmatrix} 1 & 2 & 1 & 2 \\ 0 & 2 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

Now do row operations till you can see the answer.

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

Thus $x = 1, y = 1, z = -1$.

Example 18.2.16 Find the solution to $\begin{pmatrix} 1 & 1 & 1 \\ -1 & 0 & -3 \\ 1 & 2 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 3 \end{pmatrix}$.

Add in the last column

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ -1 & 0 & -3 & 0 \\ 1 & 2 & -1 & 3 \end{pmatrix}$$

Now do row operations till you see the answer. Knowing when to stop is discussed more a little later.

$$\begin{pmatrix} 1 & 0 & 3 & 0 \\ 0 & 1 & -2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

The original system of equations has the same solution as one which includes the equation $0x + 0y + 0z = 1$ so there is no solution to this system of equations. The equations are inconsistent.

Definition 18.2.17 Let $A = (\mathbf{a}_1 \ \cdots \ \mathbf{a}_n)$. Then \mathbf{a}_k is linearly related to the other columns means there are numbers x_i such that $\mathbf{a}_k = \sum_{i \neq k} x_i \mathbf{a}_i$.

This is just a more general notion than finding the solution to a system of equations in which you obtain a linear combination of columns of A equal to \mathbf{b} in $A\mathbf{x} = \mathbf{b}$. All that is happening here is to note that there is nothing sacred about the last column in $(A|\mathbf{b})$. You can ask the same question about all the other columns, whether they are a linear combination of the other columns. It turns out that row operations preserve all linear relations.

Lemma 18.2.18 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, \dots, i_r columns of B if and only if the k^{th} column of A is a linear combination of the i_1, \dots, i_r columns of A . Furthermore, the scalars in the linear combination are the same. (The linear relationship between the k^{th} column of A and the i_1, \dots, i_r columns of A is the same as the linear relationship between the k^{th} column of B and the i_1, \dots, i_r columns of B .)

Proof: Let A equal the following matrix in which the \mathbf{a}_k are the columns

$$(\mathbf{a}_1 \ \mathbf{a}_2 \ \cdots \ \mathbf{a}_n)$$

and let B equal the following matrix in which the columns are given by the \mathbf{b}_k

$$(\mathbf{b}_1 \ \mathbf{b}_2 \ \cdots \ \mathbf{b}_n)$$

Then by Theorem 18.2.11 on Page 400 $\mathbf{b}_k = E\mathbf{a}_k$ where E is an elementary matrix. Suppose then that one of the columns of A \mathbf{a}_k is a linear combination of some other columns of A . Say $\mathbf{a}_k = \sum_{i=1}^r c_i \mathbf{a}_{i_k}$. Then multiplying by E , $\mathbf{b}_k = E\mathbf{a}_k = \sum_{i=1}^r c_i E\mathbf{a}_{i_k} = \sum_{i=1}^r c_i \mathbf{b}_{i_k}$. ■

How do you know when to stop doing row operations in solving a system of equations? This involves the row reduced echelon form.

Definition 18.2.19 Let \mathbf{e}_i denote the column vector which has all zero entries except for the i^{th} slot which is one. 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 you have encountered $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_k$, the next column is either \mathbf{e}_{k+1} or is a linear combination of the vectors, $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_k$.

Theorem 18.2.20 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 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 e_1 will refer to \mathbb{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 zero out every entry above it in the rows of A_1 . 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. ■

The process of doing this is completely routine and involves elementary school arithmetic and being careful. I have found that you are less likely to make a mistake if you do it on a blackboard and erase and replace as you go. Here is an example.

Example 18.2.21 Find the row reduced echelon form for the matrix

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

I will switch the first two rows because I don't like to work with fractions. This yields

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

Now take -3 times the top row and add to the second followed by -1 times the top row added to the bottom.

$$\begin{pmatrix} 1 & -1 & 3 & 2 \\ 0 & 5 & -8 & -5 \\ 0 & 5 & 0 & 0 \end{pmatrix}$$

Now take -1 times the second row and add to the bottom.

$$\begin{pmatrix} 1 & -1 & 3 & 2 \\ 0 & 5 & -8 & -5 \\ 0 & 0 & 8 & 5 \end{pmatrix}$$

Add the bottom to the second

$$\begin{pmatrix} 1 & -1 & 3 & 2 \\ 0 & 5 & 0 & 0 \\ 0 & 0 & 8 & 5 \end{pmatrix}$$

Then take $-1/5$ times the bottom and add to the top.

$$\begin{pmatrix} 1 & -1 & \frac{7}{5} & 1 \\ 0 & 5 & 0 & 0 \\ 0 & 0 & 8 & 5 \end{pmatrix}$$

then take $-1/5$ times the middle and add to top. Finally divide each row by the numbers down the diagonal and then take 2 times the middle and add to top then $-7/5$ times the bottom and add to top.

$$\begin{pmatrix} 1 & 0 & 0 & \frac{1}{8} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \frac{5}{8} \end{pmatrix}$$

Finally you get the row reduced echelon form for this matrix.

18.2.1 Using MATLAB

It may seem tedious to find the row reduced echelon form. You can let MATLAB do it for you. Here is an example.

```
>> A=[1 2 3 4;2 3 -11 12;3 5 6 7];
rref(A)
ans =
1.0000 0 0 -7.9286
0 1.0000 0 6.9286
0 0 1.0000 -0.6429
```

At the >> I entered the matrix $\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & -11 & 12 \\ 3 & 5 & 6 & 7 \end{pmatrix}$. The semicolon indicates you

start a new row. Then press shift enter to get to the next line and press enter and it produces the row reduced echelon form. It does it in terms of decimals.

$$\begin{pmatrix} 1.0 & 0 & 0 & -7.9286 \\ 0 & 1.0 & 0 & 6.9286 \\ 0 & 0 & 1.0 & -0.64286 \end{pmatrix}$$

If you want it in terms of fractions, you do the following.

```
>>A=[1 2 3 4;2 3 -11 12;3 5 6 7];
rref(sym(A))
ans =
[ 1, 0, 0, -111/14]
[ 0, 1, 0, 97/14]
[ 0, 0, 1, -9/14]
```

You need to have the symbolic toolbox installed with MATLAB for this option.

18.2.2 Uniqueness

I keep referring to **the** row reduced echelon form. Is there only one? This would be surprising given the infinitely many ways of doing row operations. However, it is in fact the case. Any two sequences of row operations which yield a matrix in row reduced echelon form give the same thing.

Corollary 18.2.22 *The row reduced echelon form is unique. That is if B, C are two matrices in row reduced echelon form and both are row equivalent to A , then $B = C$.*

Proof: Suppose B and C are both row reduced echelon forms for the matrix A . Then they clearly have the same zero columns since row operations leave zero columns unchanged. In reading from left to right in B , suppose e_1, \dots, e_r occur first in positions i_1, \dots, i_r respectively. The description of the row reduced echelon form means that each of these columns is not a linear combination of the preceding columns. Therefore, by Lemma 18.2.18, the same is true of the columns in positions i_1, i_2, \dots, i_r for C . It follows from the description of the row reduced echelon form that in C , e_1, \dots, e_r occur first in positions i_1, i_2, \dots, i_r . Therefore, both B and C have the sequence e_1, e_2, \dots, e_r occurring first in the positions, i_1, i_2, \dots, i_r . By Lemma 18.2.18, 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, \dots, i_k position. Also the columns after the i_r position are linear combinations of the columns in the i_1, \dots, i_r positions involving the same scalars in both matrices. This is equivalent to the assertion that each of these columns is identical and this proves the corollary. ■

Definition 18.2.23 If A is an $n \times n$ matrix, and e_1, \dots, e_r occur for the first time when viewed from left to right in positions i_1, \dots, i_r , then columns i_1, \dots, i_r in the original matrix A are called pivot columns. The rank of the matrix A is the number of these pivot columns.

From the description of the row reduced echelon form, every column in this matrix is a linear combination of the pivot columns. Therefore, from Lemma 18.2.18 the same is true for the columns of the original matrix A .

Example 18.2.24 Let $A = \begin{pmatrix} 1 & 2 & 3 & 0 \\ 1 & 2 & 3 & 2 \\ -3 & 2 & -1 & 4 \end{pmatrix}$. Identify the pivot columns and rank.

The row reduced echelon form is $\begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$ and so the pivot columns are the first, second, and last. Thus every column of A is a linear combination of these pivot columns.

18.2.3 The Inverse

Definition 18.2.25 Let A be an $n \times n$ matrix. It is said to be invertible if there is a matrix B such that $AB = BA = I$. Then B is called the inverse of A and is denoted by A^{-1} .

First of all, the inverse, if it exists, is unique. To see this suppose both B, \hat{B} work in the above definition. Then

$$\hat{B} = \hat{B}I = \hat{B}(AB) = (\hat{B}A)B = IB = B$$

This means that to show something is the inverse, it suffices to show that it acts like the inverse. If it walks like a duck and quacks like a duck, then it is a duck. However, although there are many ducks, a given matrix has at most one inverse.

Recall the elementary matrices, how if E is one of them, there is another elementary matrix of the same sort \hat{E} such that $\hat{E}E = E\hat{E} = I$. This was Theorem 18.2.11 above. Thus $\hat{E} = E^{-1}$.

Lemma 18.2.26 *A product of elementary matrices $E_1 E_2 \cdots E_n$ has an inverse and its inverse is $\hat{E}_n \hat{E}_{n-1} \cdots \hat{E}_1$, the product of the inverses in the reverse order.*

Proof: In case $n = 1$, this was shown above. Suppose it is true for n matrices. Then

$$\begin{aligned} & E_1 E_2 \cdots E_n E_{n+1} \hat{E}_{n+1} \hat{E}_n \hat{E}_{n-1} \cdots \hat{E}_1 \\ &= E_1 E_2 \cdots E_n (E_{n+1} \hat{E}_{n+1}) \hat{E}_n \hat{E}_{n-1} \cdots \hat{E}_1 \\ &= E_1 E_2 \cdots E_n I \hat{E}_n \hat{E}_{n-1} \cdots \hat{E}_1 = E_1 E_2 \cdots E_n \hat{E}_n \hat{E}_{n-1} \cdots \hat{E}_1 \end{aligned}$$

and this is I by induction. It is exactly similar in the other order.

$$\begin{aligned} & \hat{E}_{n+1} \hat{E}_n \hat{E}_{n-1} \cdots \hat{E}_1 E_1 E_2 \cdots E_n E_{n+1} \\ &= \hat{E}_{n+1} \hat{E}_n \hat{E}_{n-1} \cdots (\hat{E}_1 E_1) E_2 \cdots E_n E_{n+1} \\ &= \hat{E}_{n+1} \hat{E}_n \hat{E}_{n-1} \cdots \hat{E}_2 E_2 \cdots E_n E_{n+1} = I \end{aligned}$$

by induction, since there are now only n matrices in each of the two products. ■

Now here is the main result about inverses.

Theorem 18.2.27 *Let A be an $n \times n$ matrix. Then A is invertible if and only if the row reduced echelon form of A is I . In this case A equals a finite product of elementary matrices.*

Proof: \Leftarrow Suppose the row reduced echelon form of A is I . Then, as shown above, there are elementary matrices E_1, \dots, E_m such that $E_1 \cdots E_m A = I$. Then, by Lemma 18.2.26, $A = \hat{E}_m \cdots \hat{E}_1 I = \hat{E}_m \cdots \hat{E}_1$ and so A is the product of elementary matrices. By Lemma 18.2.26 again, A^{-1} exists and equals $E_1 \cdots E_m$.

\Rightarrow Suppose now that A is invertible. Either every column of A is a pivot column, in which case the row reduced echelon form of A , called R , is the identity or else some column is not a pivot column and in this case, R has a bottom row of zeros. I need to rule out this case. However, since the bottom row of R is all zeros, there is no solution \mathbf{x} to $R\mathbf{x} = \mathbf{e}_n$. Say $E_1 \cdots E_m A = R$ where the E_i are elementary matrices corresponding to row operations which produced R . Then $A = \hat{E}_m \cdots \hat{E}_1 R$ and so there is no solution to $A\mathbf{x} = \hat{E}_m \cdots \hat{E}_1 R\mathbf{x} = \hat{E}_m \cdots \hat{E}_1 \mathbf{b} \equiv \mathbf{c}$ because by Proposition 18.2.12, multiplying both sides of an equation by an elementary matrix preserves the solution set. Now this is a contradiction because if A^{-1} exists, then you would get a unique solution to $A\mathbf{x} = \mathbf{c}$, namely $\mathbf{x} = A^{-1}\mathbf{c}$ so the first case must hold that the row reduced echelon form of A is I . ■

Now it is not hard to give a simple algorithm for finding the inverse of an $n \times n$ matrix when it exists and to determine that there is no inverse in case it does not exist. From the above, there are elementary matrices E_i such that

$$E_1 \cdots E_m A = R$$

where R is in row reduced echelon form. If $R \neq I$, then there is no inverse. If $R = I$, then the inverse of A is $E_1 \cdots E_m = E_1 \cdots E_m I$. Thus you do a sequence of row operations to I which gives the inverse of A with the same sequence of operations applied to A yielding I . This is summarized in the procedure for finding the inverse.

Procedure 18.2.28 *Let A be an $n \times n$ matrix. Write $(A|I)$. Then do row operations until you get the row reduced echelon form. If you get I on the left, then what remains on the right will be the inverse of A . If you have a row of zeros on the left so the row reduced echelon form of A is not I , then A^{-1} does not exist.*

Example 18.2.29 Find A^{-1} where $A = \begin{pmatrix} 4 & 1 & 1 \\ 0 & 2 & -1 \\ 1 & -1 & 1 \end{pmatrix}$.

Write

$$\begin{pmatrix} 4 & 1 & 1 & 1 & 0 & 0 \\ 0 & 2 & -1 & 0 & 1 & 0 \\ 1 & -1 & 1 & 0 & 0 & 1 \end{pmatrix}$$

and do row operations to find the row reduced echelon form. This yields

$$\begin{pmatrix} 1 & 0 & 0 & 1 & -2 & -3 \\ 0 & 1 & 0 & -1 & 3 & 4 \\ 0 & 0 & 1 & -2 & 5 & 8 \end{pmatrix}$$

Now the inverse is what is on the right.

$$A^{-1} = \begin{pmatrix} 1 & -2 & -3 \\ -1 & 3 & 4 \\ -2 & 5 & 8 \end{pmatrix}$$

You should always check your work.

$$\begin{pmatrix} 1 & -2 & -3 \\ -1 & 3 & 4 \\ -2 & 5 & 8 \end{pmatrix} \begin{pmatrix} 4 & 1 & 1 \\ 0 & 2 & -1 \\ 1 & -1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

By the above discussion involving elementary matrices, a matrix obtained in this way which acts like the inverse on the left will also act like the inverse on the right so it suffices to check multiplication on only one side.

18.2.4 MATLAB and Matrix Arithmetic

To find the inverse of a square matrix in matlab, you open it and type the following. The `>>` will already be there. To enter a matrix, you list the rows in order from left to right separating the entries with commas or simply leaving a space. Then to start a new row, you enter `;` a semicolon.

`>>inv([1,2,3;5,2,7;8,2,1])` Then press enter and it will give the following:

```
ans =
-0.1667 0.0556 0.1111
0.7083 -0.3194 0.1111
-0.0833 0.1944 -0.1111
```

Note how it computed the inverse in decimals. If you want the answer in terms of fractions, you should have the symbolic toolbox installed and then you do the following:

`>>inv(sym([1,2,3;5,2,7;8,2,1]))` Then press enter and it will give the following:

```
ans =
[ -1/6, 1/18, 1/9]
[ 17/24, -23/72, 1/9]
[ -1/12, 7/36, -1/9]
```

You can do other things as well. Say you have

```
>>A=[1,2,3;5,2,7;8,2,1];B=[3,2,-5;3,11,2;-3,-1,5];
```


`C=[1,2;4,-3;7,3];D=[1,2,3;-3,2,1];`

This defines some matrices. Then suppose you wanted to find $(A^{-1}D^T + BC)^T$. You would then type

`transpose(inv(sym(A))*transpose(D)+B*C) or (inv(sym(A))*D'+B*C)'`

and press enter. This gives

`ans =`

`[-427/18, 4421/72, 1007/36]`

`[-257/18, -1703/72, 451/36]`

In matlab, A' means \bar{A}^T the conjugate transpose of A . Since everything is real here, this reduces to the transpose. Also, when entering a row in a matrix, it suffices to leave a space between the entries, but you need `;` to start a new row.

To get to a new line in MATLAB, you need to press shift enter. Notice how a `;` was placed after the definition of A, B, C, D . This tells MATLAB that you have defined something but not to say anything about it. If you don't do this, then when you press return, it will list the matrices and you don't want to see that. You just want the answer. When you have done a computation in MATLAB, you ought to go to `>>` and type "clear all" and then enter. That way, you can use the symbols again with different definition. If you don't do the "clear all" thing, it will go on thinking that A is what you defined earlier.

18.3 Exercises

1. Here are some matrices:

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 7 \end{pmatrix}, B = \begin{pmatrix} 3 & -1 & 2 \\ -3 & 2 & 1 \end{pmatrix},$$

$$C = \begin{pmatrix} 1 & 2 \\ 3 & 1 \end{pmatrix}, D = \begin{pmatrix} -1 & 2 \\ 2 & -3 \end{pmatrix}, E = \begin{pmatrix} 2 \\ 3 \end{pmatrix}.$$

Find if possible $-3A, 3B - A, AC, CB, AE, EA$. If it is not possible explain why.

2. Here are some matrices:

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 2 \\ 1 & -1 \end{pmatrix}, B = \begin{pmatrix} 2 & -5 & 2 \\ -3 & 2 & 1 \end{pmatrix},$$

$$C = \begin{pmatrix} 1 & 2 \\ 5 & 0 \end{pmatrix}, D = \begin{pmatrix} -1 & 1 \\ 4 & -3 \end{pmatrix}, E = \begin{pmatrix} 1 \\ 3 \end{pmatrix}.$$

Find if possible $-3A, 3B - A, AC, CA, AE, EA, BE, DE$. If it is not possible explain why.

3. Here are some matrices:

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 2 \\ 1 & -1 \end{pmatrix}, B = \begin{pmatrix} 2 & -5 & 2 \\ -3 & 2 & 1 \end{pmatrix},$$

$$C = \begin{pmatrix} 1 & 2 \\ 5 & 0 \end{pmatrix}, D = \begin{pmatrix} -1 & 1 \\ 4 & -3 \end{pmatrix}, E = \begin{pmatrix} 1 \\ 3 \end{pmatrix}.$$

Find if possible $-3A^T, 3B - A^T, AC, CA, AE, E^T B, BE, DE, EE^T, E^T E$. If it is not possible explain why.

4. Here are some matrices:

$$\begin{aligned} A &= \begin{pmatrix} 1 & 2 \\ 3 & 2 \\ 1 & -1 \end{pmatrix}, B = \begin{pmatrix} 2 & -5 & 2 \\ -3 & 2 & 1 \end{pmatrix}, \\ C &= \begin{pmatrix} 1 & 2 \\ 5 & 0 \end{pmatrix}, D = \begin{pmatrix} -1 \\ 4 \end{pmatrix}, E = \begin{pmatrix} 1 \\ 3 \end{pmatrix}. \end{aligned}$$

Find the following if possible and explain why it is not possible if this is the case.

$$AD, DA, D^T B, D^T BE, E^T D, DE^T.$$

5. Suppose A and B are square matrices of the same size. Which of the following are correct?

- (a) $(A - B)^2 = A^2 - 2AB + B^2$
- (b) $(AB)^2 = A^2 B^2$
- (c) $(A + B)^2 = A^2 + 2AB + B^2$
- (d) $(A + B)^2 = A^2 + AB + BA + B^2$
- (e) $A^2 B^2 = A(AB)B$
- (f) $(A + B)^3 = A^3 + 3A^2 B + 3AB^2 + B^3$
- (g) $(A + B)(A - B) = A^2 - B^2$

6. Let $A = \begin{pmatrix} -1 & -1 \\ 3 & 3 \end{pmatrix}$. Find all 2×2 matrices, B such that $AB = 0$.

7. Let $\mathbf{x} = (-1, -1, 1)$ and $\mathbf{y} = (0, 1, 2)$. Find $\mathbf{x}^T \mathbf{y}$ and $\mathbf{x} \mathbf{y}^T$ if possible.

8. Let $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}, B = \begin{pmatrix} 1 & 2 \\ 3 & k \end{pmatrix}$. Is it possible to choose k such that $AB = BA$?
If so, what should k equal?

9. Let $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}, B = \begin{pmatrix} 1 & 2 \\ 1 & k \end{pmatrix}$. Is it possible to choose k such that $AB = BA$?
If so, what should k equal?

10. Let A be an $n \times n$ matrix. Show A equals the sum of a symmetric and a skew symmetric matrix. (M is skew symmetric if $M = -M^T$. M is symmetric if $M^T = M$.)
Hint: Show that $\frac{1}{2}(A^T + A)$ is symmetric and then consider using this as one of the matrices.

11. Show every skew symmetric matrix has all zeros down the main diagonal. The main diagonal consists of every entry of the matrix which is of the form a_{ii} . It runs from the upper left down to the lower right.

12. Suppose M is a 3×3 skew symmetric matrix. Show there exists a vector $\boldsymbol{\Omega}$ such that for all $\mathbf{u} \in \mathbb{R}^3$ $M\mathbf{u} = \boldsymbol{\Omega} \times \mathbf{u}$. **Hint:** Explain why, since M is skew symmetric it is of the form

$$M = \begin{pmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{pmatrix}$$

where the ω_i are numbers. Then consider $\omega_1 \mathbf{i} + \omega_2 \mathbf{j} + \omega_3 \mathbf{k}$.

13. Using only the properties 18.2 - 18.9 show $-A$ is unique.
14. Using only the properties 18.2 - 18.9 show 0 is unique.
15. Using only the properties 18.2 - 18.9 show $0A = 0$. Here the 0 on the left is the scalar 0 and the 0 on the right is the zero for $m \times n$ matrices.
16. Using only the properties 18.2 - 18.9 and previous problems show $(-1)A = -A$.
17. Prove 18.14.
18. Prove that $I_m A = A$ where A is an $m \times n$ matrix.
19. Give an example of matrices, A, B, C such that $B \neq C$, $A \neq 0$, and yet $AB = AC$.
20. Suppose $AB = AC$ and A is an invertible $n \times n$ matrix. Does it follow that $B = C$? Explain why or why not. What if A were a non invertible $n \times n$ matrix?
21. Find your own examples:
 - (a) 2×2 matrices, A and B such that $A \neq 0, B \neq 0$ with $AB \neq BA$.
 - (b) 2×2 matrices, A and B such that $A \neq 0, B \neq 0$, but $AB = 0$.
 - (c) 2×2 matrices, A, D , and C such that $A \neq 0, C \neq D$, but $AC = AD$.
22. Give an example of a matrix A such that $A^2 = I$ and yet $A \neq I$ and $A \neq -I$.
23. Give an example of matrices, A, B such that neither A nor B equals zero and yet $AB = 0$.
24. Let $A = \begin{pmatrix} 2 & 1 \\ -1 & 3 \end{pmatrix}$. Find A^{-1} if possible. If A^{-1} does not exist, determine why.
25. Let $A = \begin{pmatrix} 0 & 1 \\ 5 & 3 \end{pmatrix}$. Find A^{-1} if possible. If A^{-1} does not exist, determine why.
26. Let $A = \begin{pmatrix} 2 & 1 \\ 3 & 0 \end{pmatrix}$. Find A^{-1} if possible. If A^{-1} does not exist, determine why.
27. Let $A = \begin{pmatrix} 2 & 1 \\ 4 & 2 \end{pmatrix}$. Find A^{-1} if possible. If A^{-1} does not exist, determine why.
28. Let A be a 2×2 matrix which has an inverse. Say $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Find a formula for A^{-1} in terms of a, b, c, d .
29. Let

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 1 & 0 & 2 \end{pmatrix}.$$

Find A^{-1} if possible. If A^{-1} does not exist, determine why.

30. Let

$$A = \begin{pmatrix} 1 & 0 & 3 \\ 2 & 3 & 4 \\ 1 & 0 & 2 \end{pmatrix}.$$

Find A^{-1} if possible. If A^{-1} does not exist, determine why.

31. Let

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 4 & 5 & 10 \end{pmatrix}.$$

Find A^{-1} if possible. If A^{-1} does not exist, determine why.

32. Let

$$A = \begin{pmatrix} 1 & 2 & 0 & 2 \\ 1 & 1 & 2 & 0 \\ 2 & 1 & -3 & 2 \\ 1 & 2 & 1 & 2 \end{pmatrix}$$

Find A^{-1} if possible. If A^{-1} does not exist, determine why.

33. Write $\begin{pmatrix} x_1 - x_2 + 2x_3 \\ 2x_3 + x_1 \\ 3x_3 \\ 3x_4 + 3x_2 + x_1 \end{pmatrix}$ in the form $A \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}$ where A is an appropriate matrix.

34. Write $\begin{pmatrix} x_1 + 3x_2 + 2x_3 \\ 2x_3 + x_1 \\ 6x_3 \\ x_4 + 3x_2 + x_1 \end{pmatrix}$ in the form $A \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}$ where A is an appropriate matrix.

35. Write $\begin{pmatrix} x_1 + x_2 + x_3 \\ 2x_3 + x_1 + x_2 \\ x_3 - x_1 \\ 3x_4 + x_1 \end{pmatrix}$ in the form $A \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}$ where A is an appropriate matrix.

36. Using the inverse of the matrix, find the solution to the systems

$$\begin{pmatrix} 1 & 0 & 3 \\ 2 & 3 & 4 \\ 1 & 0 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 3 \\ 2 & 3 & 4 \\ 1 & 0 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 0 & 3 \\ 2 & 3 & 4 \\ 1 & 0 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 3 \\ 2 & 3 & 4 \\ 1 & 0 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 3 \\ -1 \\ -2 \end{pmatrix}.$$

Now give the solution in terms of a, b , and c to

$$\begin{pmatrix} 1 & 0 & 3 \\ 2 & 3 & 4 \\ 1 & 0 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} a \\ b \\ c \end{pmatrix}.$$

37. Using the inverse of the matrix, find the solution to the system

$$\begin{pmatrix} 3 & -2 & -1 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & -1 & -1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix}.$$

38. Show that if A is an $n \times n$ invertible matrix and \mathbf{x} is a $n \times 1$ matrix such that $A\mathbf{x} = \mathbf{b}$ for \mathbf{b} an $n \times 1$ matrix, then $\mathbf{x} = A^{-1}\mathbf{b}$.
39. Prove that if A^{-1} exists and $A\mathbf{x} = \mathbf{0}$ then $\mathbf{x} = \mathbf{0}$.
40. Show that if A^{-1} exists for an $n \times n$ matrix, then it is unique. That is, if $BA = I$ and $AB = I$, then $B = A^{-1}$.
41. Show that if A is an invertible $n \times n$ matrix, then so is A^T and $(A^T)^{-1} = (A^{-1})^T$.
42. Show $(AB)^{-1} = B^{-1}A^{-1}$ by verifying that $AB(B^{-1}A^{-1}) = I$ and $B^{-1}A^{-1}(AB) = I$. **Hint:** Use Problem 40.
43. Show that $(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$ by verifying that $(ABC)(C^{-1}B^{-1}A^{-1}) = I$ and $(C^{-1}B^{-1}A^{-1})(ABC) = I$. **Hint:** Use Problem 40.
44. If A is invertible, show $(A^2)^{-1} = (A^{-1})^2$. **Hint:** Use Problem 40.
45. If A is invertible, show $(A^{-1})^{-1} = A$. **Hint:** Use Problem 40.
46. Let A and be a real $m \times n$ matrix and let $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$. Show $(A\mathbf{x}, \mathbf{y})_{\mathbb{R}^m} = (\mathbf{x}, A^T\mathbf{y})_{\mathbb{R}^n}$ where $(\cdot, \cdot)_{\mathbb{R}^k}$ denotes the dot product in \mathbb{R}^k . In the notation above, this would be written as $A\mathbf{x} \cdot \mathbf{y} = \mathbf{x} \cdot A^T\mathbf{y}$. Use the definition of matrix multiplication to do this.
47. Use the result of Problem 46 to verify directly that $(AB)^T = B^T A^T$ without making any reference to subscripts.
48. Suppose A is an $n \times n$ matrix and for each j ,

$$\sum_{i=1}^n |A_{ij}| < 1$$

Show that the infinite series $\sum_{k=0}^{\infty} A^k$ converges in the sense that the ij^{th} entry of the partial sums converge for each ij . **Hint:** Let $R \equiv \max_j \sum_{i=1}^n |A_{ij}|$. Thus $R < 1$. Show that

$$\left| \sum_i (A^2)_{ij} \right| \leq R^2.$$

Then generalize to show that $\left| \sum_i (A^m)_{ij} \right| \leq R^m$. Use this to show that the ij^{th} entry of the partial sums is a Cauchy sequence. From calculus, these converge by completeness of the real or complex numbers. Next show that $(I - A)^{-1} = \sum_{k=0}^{\infty} A^k$. The Leontief model in economics involves solving an equation for \mathbf{x} of the form

$$\mathbf{x} = A\mathbf{x} + \mathbf{b}, \text{ or } (I - A)\mathbf{x} = \mathbf{b}$$

The vector $A\mathbf{x}$ is called the intermediate demand and the vectors $A^k\mathbf{x}$ have economic meaning. From the above,

$$\mathbf{x} = I\mathbf{b} + A\mathbf{b} + A^2\mathbf{b} + \cdots$$

The series is also called the Neuman series. It is important in functional analysis.

49. Let \mathbf{a} be a fixed vector. The function $T_{\mathbf{a}}$ defined by $T_{\mathbf{a}}\mathbf{v} = \mathbf{a} + \mathbf{v}$ has the effect of translating all vectors by adding \mathbf{a} . Show this is not a linear transformation. Explain why it is not possible to realize $T_{\mathbf{a}}$ in \mathbb{R}^3 by multiplying by a 3×3 matrix.
50. In spite of Problem 49 we can represent both linear transformations and translations by matrix multiplication at the expense of using higher dimensions. This is done by the homogeneous coordinates. I will illustrate in \mathbb{R}^3 where most interest in this is found. For each vector $\mathbf{v} = (v_1, v_2, v_3)^T$, consider the vector in \mathbb{R}^4 $(v_1, v_2, v_3, 1)^T$. What happens when you do

$$\begin{pmatrix} 1 & 0 & 0 & a_1 \\ 0 & 1 & 0 & a_2 \\ 0 & 0 & 1 & a_3 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ 1 \end{pmatrix} ?$$

Describe how to consider both linear transformations and translations all at once by forming appropriate 4×4 matrices.

18.4 Subspaces Spans and Bases

The span of some vectors consists of all linear combinations of these vectors. As explained earlier, a linear combination of vectors is just a finite sum of scalars times vectors.

Definition 18.4.1 Let $\{\mathbf{u}_1, \dots, \mathbf{u}_p\}$ be some vectors in \mathbb{F}^n . A linear combination of these vectors is a sum of the following form:

$$\sum_{k=1}^p a_k \mathbf{u}_k$$

That is, it is a sum of scalars times the vectors for some choice of scalars a_1, \dots, a_p . $\text{span}(\mathbf{u}_1, \dots, \mathbf{u}_p)$ denotes the set of all linear combinations of these vectors.

Observation 18.4.2 Let $\{\mathbf{u}_1, \dots, \mathbf{u}_p\}$ be vectors in \mathbb{F}^n . Form the $n \times p$ matrix $A \equiv (\mathbf{u}_1 \ \cdots \ \mathbf{u}_p)$ which has these vectors as columns. Then

$$\text{span}(\mathbf{u}_1, \dots, \mathbf{u}_p)$$

consists of all vectors which are of the form

$$A\mathbf{x} \text{ for } \mathbf{x} \in \mathbb{F}^p.$$

Recall why this is so. A typical thing in what was just described is

$$(\mathbf{u}_1 \ \cdots \ \mathbf{u}_p) \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} = x_1\mathbf{u}_1 + \cdots + x_p\mathbf{u}_p$$

In other words, a typical vector of the form $A\mathbf{x}$ is a linear combination of the columns of A . Thus we can write either $\text{span}(\mathbf{u}_1, \dots, \mathbf{u}_p)$ or all $A\mathbf{x}$ for $\mathbf{x} \in \mathbb{F}^p$ to denote the same thing.

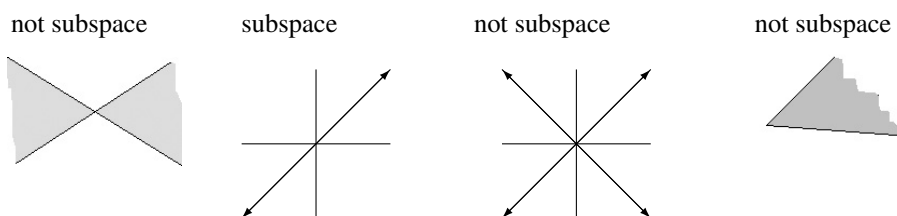
Definition 18.4.3 The vectors $A\mathbf{x}$ where $\mathbf{x} \in \mathbb{F}^p$ is also called the column space of A and also $\text{Im}(A)$ meaning image of A , also denoted as $A(\mathbb{F}^n)$. Thus column space equals $\text{span}(\mathbf{u}_1, \dots, \mathbf{u}_p)$ where the \mathbf{u}_i are the columns of A .

As explained earlier, when you say there is a solution \mathbf{x} to a linear system of equations $A\mathbf{x} = \mathbf{b}$, you mean that \mathbf{b} is in the span of the columns of A . After all, if $A = (\mathbf{u}_1 \ \dots \ \mathbf{u}_p)$, you are looking for $\mathbf{x} = (x_1 \ \dots \ x_p)^T$ such that $x_1\mathbf{u}_1 + x_2\mathbf{u}_2 + \dots + x_p\mathbf{u}_p = A\mathbf{x} = \mathbf{b}$.

A subspace is a set of vectors with the property that linear combinations of these vectors remain in the set. Geometrically, subspaces are like lines and planes which contain the origin. More precisely, the following definition is the right way to think of this.

Definition 18.4.4 Let V be a **nonempty** collection of vectors in \mathbb{F}^n . Then V is called a subspace if whenever α, β are scalars and \mathbf{u}, \mathbf{v} are vectors in V , the linear combination $\alpha\mathbf{u} + \beta\mathbf{v}$ is also in V .

There is no substitute for the above definition or equivalent algebraic definition! However, it is sometimes helpful to look at pictures at least initially. The following are four subsets of \mathbb{R}^2 . The first is the shaded area between two lines which intersect at the origin, the second is a line through the origin, the third is the union of two lines through the origin, and the last is the region between two rays from the origin. Note that in the last, multiplication of a vector in the set by a nonnegative scalar results in a vector in the set as does the sum of two vectors in the set. However, multiplication by a negative scalar does not take a vector in the set to another in the set.



Observe how the above definition indicates that the claims posted on the picture are valid. Now here are the two main examples of subspaces.

Theorem 18.4.5 Let A be an $m \times n$ matrix. Then $\text{Im}(A)$ is a subspace of \mathbb{F}^m . Also let

$$\ker(A) \equiv N(A) \equiv \{\mathbf{x} \in \mathbb{F}^n \text{ such that } A\mathbf{x} = \mathbf{0}\}$$

Then $\ker(A)$ is a subspace of \mathbb{F}^n .

Proof: Suppose $A\mathbf{x}_i$ is in $\text{Im}(A)$ and a, b are scalars. Does it follow that $aA\mathbf{x}_1 + bA\mathbf{x}_2$ is in $\text{Im}(A)$? The answer is yes because

$$aA\mathbf{x}_1 + bA\mathbf{x}_2 = A(a\mathbf{x}_1 + b\mathbf{x}_2) \in \text{Im}(A)$$

this because of the above properties of matrix multiplication. Note that $A\mathbf{0} = \mathbf{0}$ so $\mathbf{0} \in \text{Im}(A)$ and so $\text{Im}(A) \neq \emptyset$.

Now suppose \mathbf{x}, \mathbf{y} are both in $N(A)$ and a, b are scalars. Does it follow that $a\mathbf{x} + b\mathbf{y} \in N(A)$? The answer is yes because

$$A(a\mathbf{x} + b\mathbf{y}) = aA\mathbf{x} + bA\mathbf{y} = a\mathbf{0} + b\mathbf{0} = \mathbf{0}.$$

Thus the condition is satisfied. Of course $N(A) \neq \emptyset$ because $A\mathbf{0} = \mathbf{0}$. ■

Subspaces are exactly those subsets of \mathbb{F}^n which are themselves vector spaces. Recall that a vector space is something which satisfies the vector space axioms on Page 283.

Proposition 18.4.6 *Let V be a nonempty collection of vectors in \mathbb{F}^n . Then V is a subspace if and only if V is itself a vector space having the same operations as those defined on \mathbb{F}^n .*

Proof: Suppose first that V is a subspace. It is obvious all the algebraic laws hold on V because it is a subset of \mathbb{F}^n and they hold on \mathbb{F}^n . Thus $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$ along with the other axioms. Does V contain $\mathbf{0}$? Yes because it contains $0\mathbf{u} = \mathbf{0}$. Are the operations defined on V ? That is, when you add vectors of V do you get a vector in V ? When you multiply a vector in V by a scalar, do you get a vector in V ? Yes. This is contained in the definition. Does every vector in V have an additive inverse? Yes because $-\mathbf{v} = (-1)\mathbf{v}$ which is given to be in V provided $\mathbf{v} \in V$.

Next suppose V is a vector space. Then by definition, it is closed with respect to linear combinations. Hence it is a subspace. ■

18.5 Linear Independence

Now here is a very fundamental definition.

Definition 18.5.1 *Let $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ be vectors in \mathbb{F}^p . They are independent if and only if the only solution to the system of equations*

$$\begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_r \end{pmatrix} \mathbf{x} = \mathbf{0}$$

is $\mathbf{x} = \mathbf{0}$. In other words the vectors are independent means that whenever

$$\sum_{i=1}^r x_i \mathbf{u}_i = \mathbf{0}$$

it follows that each $x_i = 0$. The set of vectors is dependent if it is not independent.

Note that any list of vectors containing the zero vector is automatically linearly dependent. Indeed, you could multiply this vector by 1 and all the others by 0. Then adding these together, you would have a linear combination of the vectors in your list which equals $\mathbf{0}$ although not all of the scalars used are 0. There is a fundamental result in the case where $m < n$. In this case, the matrix A of the linear transformation looks like the following.



Theorem 18.5.2 *Let A be an $m \times n$ matrix where $m < n$. Then $N(A)$ contains nonzero vectors.*

Proof: Since the matrix has more columns than rows, you can have at most m pivot columns. Without loss of generality, A has a nonzero column. Pick the first non-pivot column of A called \mathbf{a}_r . Then such that $\mathbf{a}_r = \sum_{i=1}^{r-1} c_i \mathbf{a}_i$. Therefore,

$$\mathbf{0} = \sum_{i=1}^{r-1} c_i \mathbf{a}_i - \mathbf{a}_r = \begin{pmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_{r-1} & \mathbf{a}_r & \cdots & \mathbf{a}_n \end{pmatrix} \begin{pmatrix} c_1 \\ \vdots \\ c_{r-1} \\ -1 \\ \vdots \\ 0 \end{pmatrix}$$

■

Note that the same conclusion occurs more generally if there is some column which is not a pivot column even in case $m \geq n$.

With this preparation, here is a major theorem.

Theorem 18.5.3 *Suppose you have vectors $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ and that this set of vectors is independent. Suppose also that there are vectors $\{\mathbf{v}_1, \dots, \mathbf{v}_s\}$ and that each \mathbf{u}_j is a linear combination of the vectors $\{\mathbf{v}_1, \dots, \mathbf{v}_s\}$. Then $r \leq s$. A little less precisely, spanning sets are at least as long as linearly independent sets.*

Proof: By assumption, $\mathbf{u}_k = \sum_{i=1}^s a_{ik} \mathbf{v}_i$ for a suitable choice of the scalars a_{ik} . Then the matrix whose ik^{th} entry is a_{ik} has more columns than rows if $s < r$. Thus there is $\mathbf{x} \neq \mathbf{0}$ such that $\sum_{k=1}^r a_{ik} x_k = 0$ for each i thanks to Theorem 18.5.2. Now

$$\sum_{k=1}^r x_k \mathbf{u}_k = \sum_{k=1}^r x_k \sum_{i=1}^s a_{ik} \mathbf{v}_i = \sum_{i=1}^s \left(\sum_{k=1}^r a_{ik} x_k \right) \mathbf{v}_i = \sum_{i=1}^s 0 \mathbf{v}_i = \mathbf{0}$$

contradicting linear independence of $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ and so you must have $s \geq r$. ■

Now is the very important idea of a basis and dimension.

Definition 18.5.4 *Let V be a subspace of \mathbb{F}^n . Then $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ is called a basis for V if each $\mathbf{u}_i \in V$ and $\text{span}(\mathbf{u}_1, \dots, \mathbf{u}_r) = V$ and $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ is linearly independent. In words, $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ spans and is independent.*

Theorem 18.5.5 *Let $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ and $\{\mathbf{v}_1, \dots, \mathbf{v}_s\}$ be bases for V . Then $s = r$.*

Proof: From Theorem 18.5.3, $r \leq s$ since $\mathbf{u}_i \in \text{span}(\mathbf{v}_1, \dots, \mathbf{v}_s)$ and $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ is independent. Then also $r \geq s$ by the same reasoning. ■

Definition 18.5.6 *Let V be a subspace of \mathbb{F}^n . Then the dimension of V is the number of vectors in a basis. This is well defined by Theorem 18.5.5.*

Observation 18.5.7 *The dimension of \mathbb{F}^n is n . This is obvious because if $\mathbf{x} \in \mathbb{F}^n$, where $\mathbf{x} = \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix}^T$, then $\mathbf{x} = \sum_{i=1}^n x_i \mathbf{e}_i$ which shows that $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ is a spanning set. However, these vectors are clearly independent because if*

$$\sum_i x_i \mathbf{e}_i = \mathbf{0},$$

then $\mathbf{0} = (x_1 \cdots x_n)^T$ and so each $x_i = 0$. Thus $\{e_1, \dots, e_n\}$ is also linearly independent.

The next lemma says that if you have a vector not in the span of a linearly independent set, then you can add it in and the resulting longer list of vectors will still be linearly independent.

Lemma 18.5.8 Suppose $v \notin \text{span}(\mathbf{u}_1, \dots, \mathbf{u}_k)$ and $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$ is linearly independent. Then $\{\mathbf{u}_1, \dots, \mathbf{u}_k, v\}$ is also linearly independent.

Proof: Suppose $\sum_{i=1}^k c_i \mathbf{u}_i + d\mathbf{v} = \mathbf{0}$. It is required to verify that each $c_i = 0$ and that $d = 0$. But if $d \neq 0$, then you can solve for v as a linear combination of the vectors, $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$, $\mathbf{v} = -\sum_{i=1}^k \left(\frac{c_i}{d}\right) \mathbf{u}_i$ contrary to assumption. Therefore, $d = 0$. But then $\sum_{i=1}^k c_i \mathbf{u}_i = \mathbf{0}$ and the linear independence of $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$ implies each $c_i = 0$ also. ■

It turns out that every nonzero subspace equals the span of linearly independent vectors. This is the content of the next theorem.

Theorem 18.5.9 V is a nonzero subspace of \mathbb{F}^n if and only if it has a basis.

Proof: Pick a nonzero vector of V , \mathbf{u}_1 . If $V = \text{span}\{\mathbf{u}_1\}$, then stop. You have found your basis. If $V \neq \text{span}(\mathbf{u}_1)$, then there exists \mathbf{u}_2 a vector of V which is not a vector in $\text{span}(\mathbf{u}_1)$. Consider $\text{span}(\mathbf{u}_1, \mathbf{u}_2)$. By Lemma 18.5.8, $\{\mathbf{u}_1, \mathbf{u}_2\}$ is linearly independent. If $V = \text{span}(\mathbf{u}_1, \mathbf{u}_2)$, stop. You have found a basis. Otherwise, pick $\mathbf{u}_3 \notin \text{span}(\mathbf{u}_1, \mathbf{u}_2)$. Continue this way until you obtain a basis. The process must stop after fewer than $n + 1$ iterations because if it didn't, then there would be a linearly independent set of more than n vectors which is impossible because there is a spanning set of n vectors from the above observation. ■

The following is a fundamental result. It includes the idea that you can enlarge a linearly independent set of vectors to obtain a basis.

Theorem 18.5.10 If V is a subspace of \mathbb{F}^n and the dimension of V is m , then $m \leq n$ and also if $\{\mathbf{u}_1, \dots, \mathbf{u}_m\}$ is an independent set of vectors of V , then this set of vectors is a basis for V . Also, if you have a linearly independent set of vectors of V , $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$ for $k \leq m = \dim(V)$, there is a linearly independent set of vectors $\{\mathbf{u}_1, \dots, \mathbf{u}_k, \mathbf{v}_{k+1}, \dots, \mathbf{v}_m\}$ which is a basis for V .

Proof: If the dimension of V is m , then it has a basis of m vectors. It follows $m \leq n$ because if not, you would have an independent set of vectors which is longer than a spanning set of vectors $\{e_1, \dots, e_n\}$ contrary to Theorem 18.5.3.

Next, if $\{\mathbf{u}_1, \dots, \mathbf{u}_m\}$ is an independent set of vectors of V , then if it fails to span V , it must be there is a vector \mathbf{w} which is not in this span. But then by Lemma 18.5.8, you could add \mathbf{w} to the list of vectors and get an independent set of $m + 1$ vectors. However, the fact that V is of dimension m means there is a spanning set having only m vectors and so this contradicts Lemma 18.5.8. Thus $\{\mathbf{u}_1, \dots, \mathbf{u}_m\}$ must be a spanning set.

Finally, if $k = m$, the vectors $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$ must span V since if not, you could add another vector which is not in this list and get an independent set which is longer than a spanning set contrary to Theorem 18.5.3. Thus assume $k < m$. The set of vectors $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$ cannot span V because if it did, the dimension of V would be k not m . Thus there is a vector \mathbf{v}_{k+1} not in this span. Then by Lemma 18.5.8, $\{\mathbf{u}_1, \dots, \mathbf{u}_k, \mathbf{v}_{k+1}\}$ is

independent. If it spans V , stop. You have your basis. Otherwise, there is a \mathbf{v}_{k+2} not in the span and so you can add it in and get an independent set $\{\mathbf{u}_1, \dots, \mathbf{u}_k, \mathbf{v}_{k+1}, \mathbf{v}_{k+2}\}$. Continue this process till it stops. It must stop since otherwise, you would be able to get an independent set of vectors larger than m which is the dimension of V , contrary to Theorem 18.5.3. ■

Definition 18.5.11 *The rank of a matrix A is the dimension of $\text{Im}(A)$ which is the same as the column space of A .*

Theorem 18.5.12 *Let A be an $n \times n$ matrix. Then A^{-1} exists if and only if the rank of A equals n .*

Proof: This follows from Theorem 18.2.27 which says that A has an inverse if and only if each column of A is a pivot column if and only if the rank of A is n . ■

18.6 Exercises

- Let $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ be vectors in \mathbb{R}^n . The parallelepiped determined by these vectors $P(\mathbf{u}_1, \dots, \mathbf{u}_n)$ is defined as

$$P(\mathbf{u}_1, \dots, \mathbf{u}_n) \equiv \left\{ \sum_{k=1}^n t_k \mathbf{u}_k : t_k \in [0, 1] \text{ for all } k \right\}.$$

Now let A be an $n \times n$ matrix. Show $\{A\mathbf{x} : \mathbf{x} \in P(\mathbf{u}_1, \dots, \mathbf{u}_n)\}$ is also a parallelepiped.

- In the context of Problem 1, draw $P(\mathbf{e}_1, \mathbf{e}_2)$ where $\mathbf{e}_1, \mathbf{e}_2$ are the standard basis vectors for \mathbb{R}^2 . Thus $\mathbf{e}_1 = (1, 0), \mathbf{e}_2 = (0, 1)$. Now suppose $E = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ where E is the elementary matrix which takes the second row and adds to the first. Draw $\{E\mathbf{x} : \mathbf{x} \in P(\mathbf{e}_1, \mathbf{e}_2)\}$. In other words, draw the result of doing E to the vectors in $P(\mathbf{e}_1, \mathbf{e}_2)$. Next draw the results of doing the other elementary matrices to $P(\mathbf{e}_1, \mathbf{e}_2)$.
- Determine which matrices are in row reduced echelon form.

(a) $\begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 7 \end{pmatrix}$

(c) $\begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 5 \\ 0 & 0 & 1 & 2 & 0 & 4 \\ 0 & 0 & 0 & 0 & 1 & 3 \end{pmatrix}$

(b) $\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{pmatrix}$

- Row reduce the following matrices to obtain the row reduced echelon form. List the pivot columns in the original matrix.

(a) $\begin{pmatrix} 1 & 2 & 0 & 3 \\ 2 & 1 & 2 & 2 \\ 1 & 1 & 0 & 3 \end{pmatrix}$

(b) $\begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & -2 \\ 3 & 0 & 0 \\ 3 & 2 & 1 \end{pmatrix}$

$$(c) \begin{pmatrix} 1 & 2 & 1 & 3 \\ -3 & 2 & 1 & 0 \\ 3 & 2 & 1 & 1 \end{pmatrix}$$

5. Find the rank of the following matrices. If the rank is r , identify r columns **in the original matrix** which have the property that every other column may be written as a linear combination of these. Also find a basis for column space of the matrices.

$$(a) \begin{pmatrix} 1 & 2 & 0 \\ 3 & 2 & 1 \\ 2 & 1 & 0 \\ 0 & 2 & 1 \end{pmatrix}$$

$$(d) \begin{pmatrix} 0 & 1 & 0 & 2 & 0 & 1 & 0 \\ 0 & 3 & 2 & 6 & 0 & 5 & 4 \\ 0 & 1 & 1 & 2 & 0 & 2 & 2 \\ 0 & 2 & 1 & 4 & 0 & 3 & 2 \end{pmatrix}$$

$$(b) \begin{pmatrix} 1 & 0 & 0 \\ 4 & 1 & 1 \\ 2 & 1 & 0 \\ 0 & 2 & 0 \end{pmatrix}$$

$$(e) \begin{pmatrix} 0 & 1 & 0 & 2 & 1 & 1 & 2 \\ 0 & 3 & 2 & 6 & 1 & 5 & 1 \\ 0 & 1 & 1 & 2 & 0 & 2 & 1 \\ 0 & 2 & 1 & 4 & 0 & 3 & 1 \end{pmatrix}$$

$$(c) \begin{pmatrix} 0 & 1 & 0 & 2 & 1 & 2 & 2 \\ 0 & 3 & 2 & 12 & 1 & 6 & 8 \\ 0 & 1 & 1 & 5 & 0 & 2 & 3 \\ 0 & 2 & 1 & 7 & 0 & 3 & 4 \end{pmatrix}$$

6. Suppose A is an $m \times n$ matrix. Explain why the rank of A is always no larger than $\min(m, n)$.
7. A matrix A is called a projection if $A^2 = A$. Here is a matrix.

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

Show that this is a projection. Show that a vector in the column space of a projection matrix is left unchanged by multiplication by A .

8. Let H denote $\text{span} \left(\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 \\ 4 \end{pmatrix}, \begin{pmatrix} 1 \\ 3 \end{pmatrix} \right)$. Find the dimension of H and determine a basis.
9. Let H denote $\text{span} \left(\begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 4 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 3 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \right)$. Find the dimension of H and determine a basis.
10. Let H denote $\text{span} \left(\begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 4 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 3 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \right)$. Find the dimension of H and determine a basis.
11. Let $M = \{ \mathbf{u} = (u_1, u_2, u_3, u_4) \in \mathbb{R}^4 : u_3 = u_1 = 0 \}$. Is M a subspace? Explain.
12. Let $M = \{ \mathbf{u} = (u_1, u_2, u_3, u_4) \in \mathbb{R}^4 : u_3 \geq u_1 \}$. Is M a subspace? Explain.

13. Let $\mathbf{w} \in \mathbb{R}^4$ and let $M = \{\mathbf{u} = (u_1, u_2, u_3, u_4) \in \mathbb{R}^4 : \mathbf{w} \cdot \mathbf{u} = 0\}$. Is M a subspace? Explain.
14. Let $M = \{\mathbf{u} = (u_1, u_2, u_3, u_4) \in \mathbb{R}^4 : u_i \geq 0 \text{ for each } i = 1, 2, 3, 4\}$. Is M a subspace? Explain.
15. Let \mathbf{w}, \mathbf{w}_1 be given vectors in \mathbb{R}^4 and define

$$M = \{\mathbf{u} = (u_1, u_2, u_3, u_4) \in \mathbb{R}^4 : \mathbf{w} \cdot \mathbf{u} = 0 \text{ and } \mathbf{w}_1 \cdot \mathbf{u} = 0\}.$$

Is M a subspace? Explain.

16. Let $M = \{\mathbf{u} = (u_1, u_2, u_3, u_4) \in \mathbb{R}^4 : |u_1| \leq 4\}$. Is M a subspace? Explain.
17. Let $M = \{\mathbf{u} = (u_1, u_2, u_3, u_4) \in \mathbb{R}^4 : \sin(u_1) = 1\}$. Is M a subspace? Explain.
18. Suppose $\{\mathbf{x}_1, \dots, \mathbf{x}_k\}$ is a set of vectors from \mathbb{F}^n . Show that $\text{span}(\mathbf{x}_1, \dots, \mathbf{x}_k)$ contains $\mathbf{0}$.
19. Prove the following theorem: If A, B are $n \times n$ matrices and if $AB = I$, then $BA = I$ and $B = A^{-1}$. **Hint:** First note that if $AB = I$, then it must be the case that A is onto. Explain why this requires $\text{span}(\text{columns of } A) = \mathbb{F}^n$. Now explain why, this requires A to be one to one. Next explain why $A(BA - I) = 0$ and why the fact that A is one to one implies $BA = I$.
20. Here are three vectors. Determine whether they are linearly independent or linearly dependent. $\begin{pmatrix} 1 & 2 & 0 \end{pmatrix}^T, \begin{pmatrix} 2 & 0 & 1 \end{pmatrix}^T, \begin{pmatrix} 3 & 0 & 0 \end{pmatrix}^T$ Make them the columns of a matrix and row reduce to determine whether they are linearly independent.
21. Here are three vectors. Determine whether they are linearly independent or linearly dependent. $\begin{pmatrix} 4 & 2 & 0 \end{pmatrix}^T, \begin{pmatrix} 2 & 2 & 1 \end{pmatrix}^T, \begin{pmatrix} 0 & 2 & 2 \end{pmatrix}^T$
22. Here are three vectors. Determine whether they are linearly independent or linearly dependent. $\begin{pmatrix} 1 & 2 & 3 \end{pmatrix}^T, \begin{pmatrix} 4 & 5 & 1 \end{pmatrix}^T, \begin{pmatrix} 3 & 1 & 0 \end{pmatrix}^T$
23. Here are four vectors. Determine if they span \mathbb{R}^3 . Are these vectors linearly independent? $\begin{pmatrix} 1 & 2 & 3 \end{pmatrix}^T, \begin{pmatrix} 4 & 3 & 3 \end{pmatrix}^T, \begin{pmatrix} 3 & 1 & 0 \end{pmatrix}^T, \begin{pmatrix} 2 & 4 & 6 \end{pmatrix}^T$
24. Here are four vectors. Determine if they span \mathbb{R}^3 . Are these vectors linearly independent? $\begin{pmatrix} 1 & 2 & 3 \end{pmatrix}^T, \begin{pmatrix} 4 & 3 & 3 \end{pmatrix}^T, \begin{pmatrix} 3 & 2 & 0 \end{pmatrix}^T, \begin{pmatrix} 2 & 4 & 6 \end{pmatrix}^T$
25. Determine if the following vectors are a basis for \mathbb{R}^3 . If they are, explain why they are and if they are not, give a reason and tell whether they span \mathbb{R}^3 . $\begin{pmatrix} 1 & 0 & 3 \end{pmatrix}^T, \begin{pmatrix} 4 & 3 & 3 \end{pmatrix}^T, \begin{pmatrix} 1 & 2 & 0 \end{pmatrix}^T, \begin{pmatrix} 2 & 4 & 0 \end{pmatrix}^T$
26. Consider the vectors of the form

$$\left\{ \begin{pmatrix} 2t + 3s \\ s - t \\ t + s \end{pmatrix} : s, t \in \mathbb{R} \right\}.$$

Is this set of vectors a subspace of \mathbb{R}^3 ? If so, explain why, give a basis for the subspace and find its dimension.

27. Consider the vectors of the form

$$\left\{ \begin{pmatrix} 2t + 3s + u \\ s - t \\ t + s \\ u \end{pmatrix} : s, t, u \in \mathbb{R} \right\}.$$

Is this set of vectors a subspace of \mathbb{R}^4 ? If so, explain why, give a basis for the subspace and find its dimension.

28. Consider the vectors of the form

$$\left\{ \begin{pmatrix} 2t + u \\ t + 3u \\ t + s + v \\ u \end{pmatrix} : s, t, u, v \in \mathbb{R} \right\}.$$

Is this set of vectors a subspace of \mathbb{R}^4 ? If so, explain why, give a basis for the subspace and find its dimension.

29. If you have 5 vectors in \mathbb{F}^5 and the vectors are linearly independent, can it always be concluded they span \mathbb{F}^5 ? Explain.
30. If you have 6 vectors in \mathbb{F}^5 , is it possible they are linearly independent? Explain.
31. Suppose A is an $m \times n$ matrix and $\{w_1, \dots, w_k\}$ is a linearly independent set of vectors in $A(\mathbb{F}^n) \subseteq \mathbb{F}^m$. Now suppose $A(z_i) = w_i$. Show $\{z_1, \dots, z_k\}$ is also independent.
32. Suppose V, W are subspaces of \mathbb{F}^n . Show $V \cap W$ defined to be all vectors which are in both V and W is a subspace also.
33. Suppose V and W both have dimension equal to 7 and they are subspaces of \mathbb{F}^{10} . What are the possibilities for the dimension of $V \cap W$? **Hint:** Remember that a linear independent set can be extended to form a basis.
34. Suppose V has dimension p and W has dimension q and they are each contained in a subspace, U which has dimension equal to n where $n > \max(p, q)$. What are the possibilities for the dimension of $V \cap W$? **Hint:** Remember that a linear independent set can be extended to form a basis.
35. If $b \neq 0$, can the solution set of $Ax = b$ be a plane through the origin? Explain.
36. Suppose a system of equations has fewer equations than variables and you have found a solution to this system of equations. Is it possible that your solution is the only one? Explain.
37. Suppose a system of linear equations has a 2×4 augmented matrix and the last column is a pivot column. Could the system of linear equations be consistent? Explain.
38. Suppose the coefficient matrix of a system of n equations with n variables has the property that every column is a pivot column. Does it follow that the system of equations must have a solution? If so, must the solution be unique? Explain.

39. Suppose there is a unique solution to a system of linear equations. What must be true of the pivot columns in the augmented matrix.
40. State whether each of the following sets of data are possible for the matrix equation $Ax = b$. If possible, describe the solution set. That is, tell whether there exists a unique solution no solution or infinitely many solutions.
- (a) A is a 5×6 matrix, $\text{rank}(A) = 4$ and $\text{rank}(A|b) = 4$. **Hint:** This says b is in the span of four of the columns. Thus the columns are not independent.
 - (b) A is a 3×4 matrix, $\text{rank}(A) = 3$ and $\text{rank}(A|b) = 2$.
 - (c) A is a 4×2 matrix, $\text{rank}(A) = 4$ and $\text{rank}(A|b) = 4$. **Hint:** This says b is in the span of the columns and the columns must be independent.
 - (d) A is a 5×5 matrix, $\text{rank}(A) = 4$ and $\text{rank}(A|b) = 5$. **Hint:** This says b is not in the span of the columns.
 - (e) A is a 4×2 matrix, $\text{rank}(A) = 2$ and $\text{rank}(A|b) = 2$.
41. Suppose A is an $m \times n$ matrix in which $m \leq n$. Suppose also that the rank of A equals m . Show that A maps \mathbb{F}^n onto \mathbb{F}^m . **Hint:** The vectors e_1, \dots, e_m occur as columns in the row reduced echelon form for A .
42. Suppose A is an $m \times n$ matrix in which $m \geq n$. Suppose also that the rank of A equals n . Show that A is one to one. **Hint:** If not, there exists a vector x such that $Ax = 0$, and this implies at least one column of A is a linear combination of the others. Show this would require the column rank to be less than n .
43. Explain why an $n \times n$ matrix A is both one to one and onto if and only if its rank is n .
44. For M a matrix, $\ker(M)$ consists of all vectors x such that $Mx = 0$. Suppose A is an $m \times n$ matrix and B is an $n \times p$ matrix. Show that

$$\dim(\ker(AB)) \leq \dim(\ker(A)) + \dim(\ker(B)).$$

Hint: Consider the subspace, $B(\mathbb{F}^p) \cap \ker(A)$ and suppose a basis for this subspace is $\{w_1, \dots, w_k\}$. Now suppose $\{u_1, \dots, u_r\}$ is a basis for $\ker(B)$. Let $\{z_1, \dots, z_k\}$ be such that $Bz_i = w_i$ and argue that

$$\ker(AB) \subseteq \text{span}(u_1, \dots, u_r, z_1, \dots, z_k).$$

Here is how you do this. Suppose $ABx = 0$. Then $Bx \in \ker(A) \cap B(\mathbb{F}^p)$ and so $Bx = \sum_{i=1}^k Bz_i$ showing that $x - \sum_{i=1}^k z_i \in \ker(B)$.

45. Explain why $Ax = 0$ always has a solution even when A^{-1} does not exist.
- (a) What can you conclude about A if the solution is unique?
 - (b) What can you conclude about A if the solution is not unique?
46. Let A be an $n \times n$ matrix and let x be a nonzero vector such that $Ax = \lambda x$ for some scalar λ . When this occurs, the vector x is called an **eigenvector** and the scalar λ is called an **eigenvalue**. It turns out that not every number is an eigenvalue. Only certain ones are. Why? **Hint:** Show that if $Ax = \lambda x$, then $(A - \lambda I)x = 0$. Explain why this shows that $(A - \lambda I)$ is not one to one and not onto.

47. Let A be an $n \times n$ matrix and consider the matrices $\{I, A, A^2, \dots, A^{n^2}\}$. Explain why there exist scalars, c_i not all zero such that $\sum_{i=1}^{n^2} c_i A^i = 0$. Then argue there exists a polynomial, $p(\lambda)$ of the form

$$\lambda^m + d_{m-1}\lambda^{m-1} + \dots + d_1\lambda + d_0$$

such that $p(A) = 0$ and if $q(\lambda)$ is another polynomial such that $q(A) = 0$, then $q(\lambda)$ is of the form $p(\lambda)l(\lambda)$ for some polynomial, $l(\lambda)$. This extra special polynomial, $p(\lambda)$ is called the **minimal polynomial**. **Hint:** You might consider an $n \times n$ matrix as a vector in \mathbb{F}^{n^2} . What would be a basis for this set of matrices?

48. Let A be an $n \times n$ matrix and let $p(\lambda)$ be the minimal polynomial of the above problem. By the fundamental theorem of algebra, this can be factored as

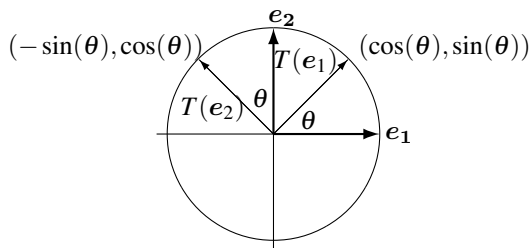
$$\prod_{i=1}^m (\lambda - \mu_i)$$

where $\mu_i \in \mathbb{C}$. Thus, from the above problem, $\prod_{i=1}^m (A - \mu_i I) = 0$. Explain why there is a vector v_k such that $u_k \equiv \prod_{i \neq k} (A - \mu_i I) v_k \neq 0$. Explain why $(A - \mu_k I) u_k = 0$. Thus A has an eigenvector for each of the μ_i . Note that you must allow all arithmetic to take place in \mathbb{C} because the eigenvalues μ_i are only known to be complex numbers.

49. Let $\theta \in \mathbb{R}$. For x a vector in \mathbb{R}^p , $p > 1$, let T_θ be defined as follows. Place x with its tail at the origin and rotate through an angle of θ . If $\theta > 0$, rotate counter clockwise and if $\theta < 0$ rotate clockwise as in trigonometry. Argue with elementary geometry that T_θ is a linear transformation. In case $p = 2$, explain why the matrix of T_θ , called a rotation matrix, is

$$\begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}$$

It amounts to justifying the following picture.



50. Now note that $T_\theta T_\alpha = T_{\theta+\alpha}$. Using matrix multiplication and the above problem, derive with virtually no effort the formulas for $\sin(\theta + \alpha)$ and $\cos(\theta + \alpha)$.

Chapter 19

Eigenvalues and Eigenvectors

19.1 Definition of Eigenvalues

The thing to always keep in mind is the following definition of eigenvalues and eigenvectors. There are many ways to find them and in this chapter, I will present the standard way to do this. It is also the very worst way. This is a book on multi-variable calculus, not one on linear algebra. This is why I have been focussed almost exclusively on \mathbb{R}^n . However, when one considers eigenvalues and eigenvectors, it is no longer possible to give a reasonable presentation without the use of the complex numbers. Thus, for the material in this section, it will be understood that the vectors are in \mathbb{C}^n meaning ordered lists of complex numbers. The matrices will also be understood to have entries in \mathbb{C} and all scalars will be understood to lie in \mathbb{C} rather than be restricted to be in \mathbb{R} .

Definition 19.1.1 *Let A be an $n \times n$ matrix and let $x \in \mathbb{C}^n, \lambda \in \mathbb{C}$. Then x is an eigenvector for the eigenvalue λ if and only if the following two conditions hold.*

1. $Ax = \lambda x$
2. $x \neq 0$. **This is very important. By definition 0 is NEVER an eigenvector although 0 can be an eigenvalue.**

Now here is an important observation which really is just a re statement of the above definition.

Theorem 19.1.2 *Let A be an $n \times n$ matrix. The vector x is an eigenvector for the eigenvalue λ if and only if $(A - \lambda I)^{-1}$ does not exist.*

Proof: If $(A - \lambda I)^{-1}$ does not exist, then by Theorem 18.5.12 the columns of $A - \lambda I$ are not independent because its rank is less than n . Thus there exists $x \neq 0$ such that $(A - \lambda I)x = 0$ and so λ is an eigenvalue and x is an eigenvector which goes with λ . Conversely, if $(A - \lambda I)x = 0$, and $x \neq 0$, then the rank of $(A - \lambda I)$ has no inverse because its rank is less than n . Indeed, some column is a linear combination of the others. ■

Now with this fundamental definition, I will present the worst way of finding eigenvalues and eigenvectors. It is very important because everyone cherishes it and it is the standard way to do it in all undergraduate courses. Also, it gives an introduction to the important topic of determinants which will be presented in more detail later.

19.2 An Introduction to Determinants

19.2.1 Cofactors and 2×2 Determinants

Let A be an $n \times n$ matrix. The **determinant** of A , denoted as $\det(A)$ is a number. If the matrix is a 2×2 matrix, this number is very easy to find.

Definition 19.2.1 Let $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Then $\det(A) \equiv ad - cb$. The determinant is also often denoted by enclosing the matrix with two vertical lines. Thus

$$\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{vmatrix} a & b \\ c & d \end{vmatrix}.$$

Example 19.2.2 Find $\det \begin{pmatrix} 2 & 4 \\ -1 & 6 \end{pmatrix}$.

From the definition this is just $(2)(6) - (-1)(4) = 16$.

Having defined what is meant by the determinant of a 2×2 matrix, what about a 3×3 matrix?

Definition 19.2.3 Suppose A is a 3×3 matrix. The ij^{th} **minor**, denoted here as $\text{minor}(A)_{ij}$, is the determinant of the 2×2 matrix which results from deleting the i^{th} row and the j^{th} column.

Example 19.2.4 Consider the matrix

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

The $(1,2)$ minor is the determinant of the 2×2 matrix which results when you delete the first row and the second column. This minor is therefore

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

The $(2,3)$ minor is the determinant of the 2×2 matrix which results when you delete the second row and the third column. This minor is therefore

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

Definition 19.2.5 Suppose A is a 3×3 matrix. The ij^{th} **cofactor** is defined to be $(-1)^{i+j} \times (ij^{\text{th}} \text{ minor})$. In words, you multiply $(-1)^{i+j}$ times the ij^{th} minor to get the ij^{th} cofactor. The cofactors of a matrix are so important that special notation is appropriate when referring to them. The ij^{th} cofactor of a matrix A will be denoted by $\text{cof}(A)_{ij}$. It is also convenient to refer to the cofactor of an entry of a matrix as follows. For a_{ij} an entry of the matrix, its cofactor is just $\text{cof}(A)_{ij}$. Thus the cofactor of the ij^{th} entry is just the ij^{th} cofactor.

Example 19.2.6 Consider the matrix

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 3 & 2 \\ 3 & 2 & 1 \end{pmatrix}.$$

The $(1,2)$ minor is the determinant of the 2×2 matrix which results when you delete the first row and the second column. This minor is therefore

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

It follows

$$\text{cof}(A)_{12} = (-1)^{1+2} \det \begin{pmatrix} 4 & 2 \\ 3 & 1 \end{pmatrix} = (-1)^{1+2} (-2) = 2$$

The $(2,3)$ minor is the determinant of the 2×2 matrix which results when you delete the second row and the third column. This minor is therefore

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

Therefore,

$$\text{cof}(A)_{23} = (-1)^{2+3} \det \begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix} = (-1)^{2+3} (-4) = 4.$$

Similarly,

$$\text{cof}(A)_{22} = (-1)^{2+2} \det \begin{pmatrix} 1 & 3 \\ 3 & 1 \end{pmatrix} = -8.$$

Definition 19.2.7 The determinant of a 3×3 matrix A , is obtained by picking a row (column) and taking the product of each entry in that row (column) with its cofactor and adding these. This process when applied to the i^{th} row (column) is known as expanding the determinant along the i^{th} row (column).

Example 19.2.8 Find the determinant of

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 3 & 2 \\ 3 & 2 & 1 \end{pmatrix}.$$

Here is how it is done by “expanding along the first column”.

$$\overbrace{1(-1)^{1+1} \begin{vmatrix} 3 & 2 \\ 2 & 1 \end{vmatrix}}^{\text{cof}(A)_{11}} + \overbrace{4(-1)^{2+1} \begin{vmatrix} 2 & 3 \\ 2 & 1 \end{vmatrix}}^{\text{cof}(A)_{21}} + \overbrace{3(-1)^{3+1} \begin{vmatrix} 2 & 3 \\ 3 & 2 \end{vmatrix}}^{\text{cof}(A)_{31}} = 0.$$

This simply follows the rule in the above definition. We took the 1 in the first column and multiplied it by its cofactor, the 4 in the first column and multiplied it by its cofactor, and the 3 in the first column and multiplied it by its cofactor. Then we added these numbers together.

You could also expand the determinant along the second row as follows.

$$\overbrace{4(-1)^{2+1}}^{\text{cof}(A)_{21}} \begin{vmatrix} 2 & 3 \\ 2 & 1 \end{vmatrix} + \overbrace{3(-1)^{2+2}}^{\text{cof}(A)_{22}} \begin{vmatrix} 1 & 3 \\ 3 & 1 \end{vmatrix} + \overbrace{2(-1)^{2+3}}^{\text{cof}(A)_{23}} \begin{vmatrix} 1 & 2 \\ 3 & 2 \end{vmatrix} = 0.$$

Observe this gives the same number. You should try expanding along other rows and columns. If you don't make any mistakes, you will always get the same answer.

What about a 4×4 matrix? You know now how to find the determinant of a 3×3 matrix. The pattern is the same. In general, it is as described in the following definition.

Definition 19.2.9 Let $A = (a_{ij})$ be an $n \times n$ matrix and suppose the determinant of a $(n-1) \times (n-1)$ matrix has been defined. Then a new matrix called the **cofactor matrix**, $\text{cof}(A)$ is defined by $\text{cof}(A)_{ij} = (c_{ij})$ 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 $(-1)^{i+j}$. Thus $(-1)^{i+j} \times (\text{the } ij^{\text{th}} \text{ minor})$ equals the ij^{th} cofactor. Then $\det(A)$ is given by $\sum_i A_{ij}c_{ij} = \sum_j A_{ij}c_{ij}$. Any of these expansions along a row or a column gives the same number.

You should regard the above claim that you always get the same answer by picking any row or column with considerable skepticism. It is incredible and not at all obvious. However, it requires a little effort to establish it. This is done in the section on the theory of the determinant, Section 20 which is presented later. This is summarized in the following theorem whose conclusion is incredible.

Theorem 19.2.10 Expanding the $n \times n$ matrix along any row or column always gives the same answer so the above definition is a good definition.

Example 19.2.11 Expand $\begin{vmatrix} 1 & 2 & -1 & 1 \\ 2 & 3 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 2 & 3 & 1 \end{vmatrix}$ along first column.

It is

$$1 \begin{vmatrix} 3 & 1 & 1 \\ 1 & 0 & 0 \\ 2 & 3 & 1 \end{vmatrix} - 2 \begin{vmatrix} 2 & -1 & 1 \\ 1 & 0 & 0 \\ 2 & 3 & 1 \end{vmatrix} + 1 \begin{vmatrix} 2 & -1 & 1 \\ 3 & 1 & 1 \\ 2 & 3 & 1 \end{vmatrix} - 1 \begin{vmatrix} 2 & -1 & 1 \\ 3 & 1 & 1 \\ 1 & 0 & 0 \end{vmatrix} = 0$$

19.2.2 The Determinant of a Triangular Matrix

Notwithstanding the difficulties involved in using the method of Laplace expansion, certain types of matrices are very easy to deal with.

Definition 19.2.12 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.

You should verify the following using the above theorem on Laplace expansion.

Corollary 19.2.13 *Let M be an upper (lower) triangular matrix. Then $\det(M)$ is obtained by taking the product of the entries on the main diagonal.*

Example 19.2.14 *Let*

$$A = \begin{pmatrix} 1 & 2 & 3 & 77 \\ 0 & 2 & 6 & 7 \\ 0 & 0 & 3 & 33.7 \\ 0 & 0 & 0 & -1 \end{pmatrix}$$

Find $\det(A)$.

From the above corollary, it suffices to take the product of the diagonal elements. Thus $\det(A) = 1 \times 2 \times 3 \times (-1) = -6$. Without using the corollary, you could expand along the first column. This gives

$$\begin{aligned} & 1 \begin{vmatrix} 2 & 6 & 7 \\ 0 & 3 & 33.7 \\ 0 & 0 & -1 \end{vmatrix} + 0(-1)^{2+1} \begin{vmatrix} 2 & 3 & 77 \\ 0 & 3 & 33.7 \\ 0 & 0 & -1 \end{vmatrix} + \\ & 0(-1)^{3+1} \begin{vmatrix} 2 & 3 & 77 \\ 2 & 6 & 7 \\ 0 & 0 & -1 \end{vmatrix} + 0(-1)^{4+1} \begin{vmatrix} 2 & 3 & 77 \\ 2 & 6 & 7 \\ 0 & 3 & 33.7 \end{vmatrix} \end{aligned}$$

and the only nonzero term in the expansion is

$$1 \begin{vmatrix} 2 & 6 & 7 \\ 0 & 3 & 33.7 \\ 0 & 0 & -1 \end{vmatrix}.$$

Now expand this along the first column to obtain

$$\begin{aligned} & 1 \times \left(2 \times \begin{vmatrix} 3 & 33.7 \\ 0 & -1 \end{vmatrix} + 0(-1)^{2+1} \begin{vmatrix} 6 & 7 \\ 0 & -1 \end{vmatrix} + 0(-1)^{3+1} \begin{vmatrix} 6 & 7 \\ 3 & 33.7 \end{vmatrix} \right) \\ & = 1 \times 2 \times \begin{vmatrix} 3 & 33.7 \\ 0 & -1 \end{vmatrix} \end{aligned}$$

Next expand this last determinant along the first column to obtain the above equals

$$1 \times 2 \times 3 \times (-1) = -6$$

which is just the product of the entries down the main diagonal of the original matrix. It works this way in general.

19.2.3 Properties of Determinants

There are many properties satisfied by determinants. Some of these properties have to do with row operations. Recall the row operations.

Definition 19.2.15 *The row operations consist of the following*

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

Theorem 19.2.16 *Let A be an $n \times n$ matrix and let A_1 be a matrix which results from multiplying some row of A by a scalar c . Then $c \det(A) = \det(A_1)$.*

Example 19.2.17 *Let $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$, $A_1 = \begin{pmatrix} 2 & 4 \\ 3 & 4 \end{pmatrix}$. $\det(A) = -2$, $\det(A_1) = -4$.*

Theorem 19.2.18 *Let A be an $n \times n$ matrix and let A_1 be a matrix which results from switching two rows of A . Then $\det(A) = -\det(A_1)$. Also, if one row of A is a multiple of another row of A , then $\det(A) = 0$.*

Example 19.2.19 *Let $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ and let $A_1 = \begin{pmatrix} 3 & 4 \\ 1 & 2 \end{pmatrix}$. $\det A = -2$, $\det(A_1) = 2$.*

Theorem 19.2.20 *Let A be an $n \times n$ matrix and let A_1 be a matrix which results from applying row operation 3. That is you replace some row by a multiple of another row added to itself. Then $\det(A) = \det(A_1)$.*

Example 19.2.21 *Let $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ and let $A_1 = \begin{pmatrix} 1 & 2 \\ 4 & 6 \end{pmatrix}$. Thus the second row of A_1 is one times the first row added to the second row. $\det(A) = -2$ and $\det(A_1) = -2$.*

Theorem 19.2.22 *In Theorems 19.2.16 - 19.2.20 you can replace the word, “row” with the word “column”.*

There are two other major properties of determinants which do not involve row operations overtly.

Theorem 19.2.23 *Let A and B be two $n \times n$ matrices. Then*

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

Also,

$$\det(A) = \det(A^T).$$

Example 19.2.24 *Compare $\det(AB)$ and $\det(A) \det(B)$ for*

$$A = \begin{pmatrix} 1 & 2 \\ -3 & 2 \end{pmatrix}, B = \begin{pmatrix} 3 & 2 \\ 4 & 1 \end{pmatrix}.$$

First

$$AB = \begin{pmatrix} 1 & 2 \\ -3 & 2 \end{pmatrix} \begin{pmatrix} 3 & 2 \\ 4 & 1 \end{pmatrix} = \begin{pmatrix} 11 & 4 \\ -1 & -4 \end{pmatrix}$$

and so

$$\det(AB) = \det \begin{pmatrix} 11 & 4 \\ -1 & -4 \end{pmatrix} = -40.$$

Now

$$\det(A) = \det \begin{pmatrix} 1 & 2 \\ -3 & 2 \end{pmatrix} = 8, \det(B) = \det \begin{pmatrix} 3 & 2 \\ 4 & 1 \end{pmatrix} = -5.$$

Thus $\det(A)\det(B) = 8 \times (-5) = -40$.

19.2.4 Finding Determinants Using Row Operations

Theorems 19.2.20 - 19.2.22 can be used to find determinants using row operations. As pointed out above, the method of Laplace expansion will not be practical for any matrix of large size. Here is an example in which all the row operations are used.

Example 19.2.25 Find the determinant of the matrix

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 1 & 2 & 3 \\ 4 & 5 & 4 & 3 \\ 2 & 2 & -4 & 5 \end{pmatrix}$$

Replace the second row by (-5) times the first row added to it. Then replace the third row by (-4) times the first row added to it. Finally, replace the fourth row by (-2) times the first row added to it. This yields the matrix

$$B = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & -9 & -13 & -17 \\ 0 & -3 & -8 & -13 \\ 0 & -2 & -10 & -3 \end{pmatrix}$$

and from Theorem 19.2.20, it has the same determinant as A . Now using other row operations, $\det(B) = \left(\frac{-1}{3}\right) \det(C)$ where

$$C = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 0 & 11 & 22 \\ 0 & -3 & -8 & -13 \\ 0 & 6 & 30 & 9 \end{pmatrix}.$$

The second row was replaced by (-3) times the third row added to the second row. By Theorem 19.2.20 this didn't change the value of the determinant. Then the last row was multiplied by (-3) . By Theorem 19.2.16 the resulting matrix has a determinant which is (-3) times the determinant of the un-multiplied matrix. Therefore, we multiplied by $-1/3$ to retain the correct value. Now replace the last row with 2 times the third added to it. This does not change the value of the determinant by Theorem 19.2.20. Finally switch the third and second rows. This causes the determinant to be multiplied by (-1) . Thus $\det(C) = -\det(D)$ where

$$D = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & -3 & -8 & -13 \\ 0 & 0 & 11 & 22 \\ 0 & 0 & 14 & -17 \end{pmatrix}$$

You could do more row operations or you could note that this can be easily expanded along the first column followed by expanding the 3×3 matrix which results along its first column. Thus

$$\det(D) = 1(-3) \begin{vmatrix} 11 & 22 \\ 14 & -17 \end{vmatrix} = 1485$$

and so $\det(C) = -1485$ and $\det(A) = \det(B) = \left(-\frac{1}{3}\right)(-1485) = 495$.

Example 19.2.26 Find the determinant of the matrix

$$\begin{pmatrix} 1 & 2 & 3 & 2 \\ 1 & -3 & 2 & 1 \\ 2 & 1 & 2 & 5 \\ 3 & -4 & 1 & 2 \end{pmatrix}$$

Replace the second row by (-1) times the first row added to it. Next take -2 times the first row and add to the third and finally take -3 times the first row and add to the last row. This yields

$$\begin{pmatrix} 1 & 2 & 3 & 2 \\ 0 & -5 & -1 & -1 \\ 0 & -3 & -4 & 1 \\ 0 & -10 & -8 & -4 \end{pmatrix}.$$

By Theorem 19.2.20 this matrix has the same determinant as the original matrix. Remember you can work with the columns also. Take -5 times the last column and add to the second column. This yields

$$\begin{pmatrix} 1 & -8 & 3 & 2 \\ 0 & 0 & -1 & -1 \\ 0 & -8 & -4 & 1 \\ 0 & 10 & -8 & -4 \end{pmatrix}$$

By Theorem 19.2.22 this matrix has the same determinant as the original matrix. Now take (-1) times the third row and add to the top row. This gives.

$$\begin{pmatrix} 1 & 0 & 7 & 1 \\ 0 & 0 & -1 & -1 \\ 0 & -8 & -4 & 1 \\ 0 & 10 & -8 & -4 \end{pmatrix}$$

which by Theorem 19.2.20 has the same determinant as the original matrix. Lets expand it now along the first column. This yields the following for the determinant of the original matrix.

$$\det \begin{pmatrix} 0 & -1 & -1 \\ -8 & -4 & 1 \\ 10 & -8 & -4 \end{pmatrix}$$

which equals

$$8 \det \begin{pmatrix} -1 & -1 \\ -8 & -4 \end{pmatrix} + 10 \det \begin{pmatrix} -1 & -1 \\ -4 & 1 \end{pmatrix} = -82$$

I suggest you do not try to be fancy in using row operations. That is, stick mostly to the one which replaces a row or column with a multiple of another row or column added to

it. Also note there is no way to check your answer other than working the problem more than one way. To be sure you have gotten it right you must do this. Unfortunately, this process can go on and on when you keep getting different answers. This is a good example of something for which you should use a computer algebra system.

19.3 MATLAB and Determinants

MATLAB can find determinants. Here is an example.

```
>> A=[1,3,2,4;-5,7,2,3;2,3,7,11;1,2,3,4]; det(A)
```

Then press enter and you get

```
ans =
```

```
-102.0000
```

To enter a complex number $1 + 2i$ for example, you type: `complex(1,2)`. However, when MATLAB gives the answer, it will write it in the usual form $1 + 2i$. If you have matrices in which there are complex entries, you can go ahead and let MATLAB do the tedious computations for you.

19.4 Applications

19.4.1 A Formula for the Inverse

The definition of the determinant in terms of Laplace expansion along a row or column also provides a way to give a formula for the inverse of a matrix. Recall the definition of the inverse of a matrix in Definition 18.2.25 on Page 406. Also recall the definition of the cofactor matrix given in Definition 19.2.9 on Page 428. This cofactor matrix was just the matrix which results from replacing the ij^{th} entry of the matrix with the ij^{th} cofactor.

The following theorem 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 . In other words, A^{-1} is equal to one divided by the determinant of A times the adjugate matrix of A . This is what the following theorem says with more precision. The proof is presented later in the appendix devoted to the theory of the determinant.

Theorem 19.4.1 A^{-1} exists if and only if $\det(A) \neq 0$. If $\det(A) \neq 0$, then $A^{-1} = (a_{ij}^{-1})$ where

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

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

Example 19.4.2 Find the inverse of the matrix

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 0 & 1 \\ 1 & 2 & 1 \end{pmatrix}$$

First find the determinant of this matrix. Using Theorems 19.2.20 - 19.2.22 on Page 430, the determinant of this matrix is 12. The cofactor matrix of A is

$$\begin{pmatrix} -2 & -2 & 6 \\ 4 & -2 & 0 \\ 2 & 8 & -6 \end{pmatrix}.$$

Each entry of A was replaced by its cofactor. Therefore, from the above theorem, the inverse of A should equal

$$\frac{1}{12} \begin{pmatrix} -2 & -2 & 6 \\ 4 & -2 & 0 \\ 2 & 8 & -6 \end{pmatrix}^T = \begin{pmatrix} -1/6 & 1/3 & 1/6 \\ -1/6 & -1/6 & 2/3 \\ 1/2 & 0 & -1/2 \end{pmatrix}.$$

Does it work? You should check to see if it does. When the matrices are multiplied

$$\begin{pmatrix} -1/6 & 1/3 & 1/6 \\ -1/6 & -1/6 & 2/3 \\ 1/2 & 0 & -1/2 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 3 & 0 & 1 \\ 1 & 2 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

and so we got it right. If the result of multiplying these matrices had been something other than the identity matrix, you would know there was an error. When this happens, you need to search for the mistake if you are interested in getting the right answer. A common mistake is to forget to take the transpose of the cofactor matrix.

This formula for the inverse is also what justifies Cramer's rule.

Procedure 19.4.3 Suppose A is an $n \times n$ matrix and it is desired to solve the system $A\mathbf{x} = \mathbf{y}$, $\mathbf{y} = (y_1, \dots, y_n)^T$ for $\mathbf{x} = (x_1, \dots, x_n)^T$. Then Cramer's rule says

$$x_i = \frac{\det A_i}{\det A}$$

where A_i is obtained from A by replacing the i^{th} column of A with the column

$$(y_1, \dots, y_n)^T.$$

Find x

$$\begin{pmatrix} 1 & 2 & 1 \\ 3 & 2 & 1 \\ 2 & -3 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}.$$

From Cramer's rule,

$$x = \det \begin{pmatrix} 1 & 2 & 1 \\ 2 & 2 & 1 \\ 3 & -3 & 2 \end{pmatrix} / \det \begin{pmatrix} 1 & 2 & 1 \\ 3 & 2 & 1 \\ 2 & -3 & 2 \end{pmatrix} = \frac{1}{2}$$

To find y, z you do something similar replacing the y or z column with the right hand side.

19.4.2 Finding Eigenvalues Using Determinants

Theorem 19.4.1 says that A^{-1} exists if and only if $\det(A) \neq 0$ when there is even a formula for the inverse. Recall also that an eigenvector for λ is a nonzero vector x such that $Ax = \lambda x$ where λ is called an eigenvalue. Thus you have $(A - \lambda I)x = 0$ for $x \neq 0$. If $(A - \lambda I)^{-1}$ were to exist, then you could multiply by it on the left and obtain $x = 0$ after all. Therefore, it must be the case that $\det(A - \lambda I) = 0$. This yields a polynomial of degree n equal to 0. This polynomial is called the **characteristic polynomial**. For example, consider

$$\begin{pmatrix} 1 & -1 & -1 \\ 0 & 3 & 2 \\ 0 & -1 & 0 \end{pmatrix}$$

You need to have

$$\det \left(\begin{pmatrix} 1 & -1 & -1 \\ 0 & 3 & 2 \\ 0 & -1 & 0 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right) = 0$$

That on the left equals a polynomial of degree 3 which when factored yields

$$(1 - \lambda)(\lambda - 1)(\lambda - 2)$$

Therefore, the possible eigenvalues are 1, 1, 2. Note how the 1 is listed twice. This is because it occurs twice as a root of the characteristic polynomial. Also, if M^{-1} does not exist where M is an $n \times n$ matrix, then this means that the columns of M cannot be linearly independent since if they were, then by Theorem 18.5.12 M^{-1} would exist. Thus if $A - \lambda I$ fails to have an inverse as above, then the columns are not independent and so there exists a nonzero x such that $(A - \lambda I)x = 0$. Thus we have the following proposition.

Proposition 19.4.4 *The eigenvalues of an $n \times n$ matrix are the roots of*

$$\det(A - \lambda I) = 0.$$

Corresponding to each of these λ is an eigenvector. Every $n \times n$ matrix for $n \geq 1$ has eigenvectors and eigenvalues in \mathbb{C}^n .

Proof: It only remains to consider the last claim. This claim follows from the fundamental theorem of algebra, Theorem 15.14.3. Indeed, the characteristic polynomial is a polynomial of degree n . It has a zero λ_1 by the fundamental theorem of calculus. Thus

$$\det(A - \lambda I) = (z - \lambda_1) p_2(z)$$

where $p_2(z)$ is a polynomial of degree $n - 1$. Now apply the fundamental theorem of algebra to this one and continue this process until you obtain an expression of the form

$$\det(A - \lambda I) = (z - \lambda_1) \cdots (z - \lambda_n) (-1)^n$$

then there are n eigenvalues with some maybe being repeated. ■

Note that if $A = S^{-1}BS$, then A, B have the same characteristic polynomial, hence the same eigenvalues. (They might have different eigenvectors and usually will.) To see this, note that from the properties of determinants

$$\begin{aligned} \det(A - \lambda I) &= \det(S^{-1}BS - \lambda S^{-1}IS) = \det(S^{-1}(B - \lambda I)S) \\ &= \det(S^{-1}) \det(B - \lambda I) \det(S) = \det(S^{-1}S) \det(B - \lambda I) \\ &= \det(I) \det(B - \lambda I) = \det(B - \lambda I) \end{aligned} \tag{19.1}$$

19.5 MATLAB and Eigenvalues

The problem with eigenvalues and eigenvectors is that you have to factor a polynomial in order to get the eigenvalues. We can't do this in general. All we can do is find the eigenvalues approximately. But an approximate eigenvalue is never enough to get the eigenvector because $(A - \lambda I)^{-1}$ will exist if λ is not exactly right.

However, there are numerical methods to do this in the case that the polynomial does not factor. I am going to mention how to get the answer using MATLAB.

To find the eigenvalues enter A and follow with $;$. Then type $\text{eig}(A)$ and press return. It will give numerical approximation of the eigenvalues. If you want to have it find the exact values, you type $\text{eig}(\text{sym}(A))$ and press return. To do this last thing, you need to have the symbolic math package installed.

For example, if your matrix is

$$\begin{pmatrix} 1 & 1 & 0 \\ -1 & 0 & -1 \\ 2 & 1 & 3 \end{pmatrix},$$

You would type the following: $>>A=[1,1,0;-1,0,-1;2,1,3];$ and then $\text{eig}(\text{sym}(A))$ and return, you will get the eigenvalues 1,1,2 listed in a column. This is correct. The matrix has a repeated eigenvalue of 1. If you want to get the eigenvectors also, you would type $>>A=[1,1,0;-1,0,-1;2,1,3];$ and then $[V,D]=\text{eig}(\text{sym}(A))$ and enter or if you want numerical answers, which will sometimes be all that is available, you would type $[V,D]=\text{eig}(A)$. It will find the matrix V such that $AV = VD$ where D is a diagonal. In the case just considered, it will only find two columns for V because this is a defective matrix. In general, however, this would give $V^{-1}AV = D$ and the columns of V are the eigenvectors.

19.6 Matrices and the Dot Product

Here I will revert to consideration of \mathbb{R}^n rather than \mathbb{C}^n . I do this because this is not a book on linear algebra, only multi-variable calculus and I will give specialized treatments of some important theorems. Recall the inner product or dot product.

$$\mathbf{a} \cdot \mathbf{b} \equiv \sum_k a_k b_k$$

In more advanced contexts, this is usually written as $\langle \mathbf{a}, \mathbf{b} \rangle$ or often simply as (\mathbf{a}, \mathbf{b}) instead of $\mathbf{a} \cdot \mathbf{b}$. Also, the term “inner product” tends to be preferred over “dot product”. I will sometimes use the notation (\mathbf{a}, \mathbf{b}) instead of $\mathbf{a} \cdot \mathbf{b}$ because of this. First is an important relationship between the inner product and the transpose.

Proposition 19.6.1 *Suppose \mathbf{a}, \mathbf{b} are vectors in \mathbb{R}^n and \mathbb{R}^m respectively and let A be an $m \times n$ matrix. Then $(A\mathbf{a}, \mathbf{b}) = (\mathbf{a}, A^T \mathbf{b})$.*

Proof: From the definition of the inner product,

$$\begin{aligned} (A\mathbf{a}, \mathbf{b}) &\equiv \sum_i (A\mathbf{a})_i b_i = \sum_i \sum_j A_{ij} a_j b_i = \sum_j \sum_i A_{ij} a_j b_i \\ &= \sum_j \sum_i A_{ji}^T b_i a_j = \sum_j (A^T \mathbf{b})_j a_j = (\mathbf{a}, A^T \mathbf{b}) \blacksquare \end{aligned}$$

In words, the above says that when you take the A across the dot or comma you put a transpose on it and everything works just fine.

There are other more elegant ways to discuss eigenvectors and eigenvalues. See my book on linear algebra and analysis to see a presentation which is independent of determinants. However, this is a book on calculus, not linear algebra and the determinant is important in other contexts. Also, from the point of view of history, the determinant came earlier than the other linear algebra concepts.

19.7 Distance and Orthogonal Matrices

Some matrices preserve lengths of vectors. That is $|Ux| = |x|$ for any x in \mathbb{R}^n . Such a matrix is called orthogonal. Actually, this is not the standard definition. The standard definition is given next. First recall that if you have two square matrices of the same size and one acts like the inverse of the other on one side, then it will act like the inverse on the other side as well. See, for example, the discussion after Theorem 18.5.12. The traditional definition of orthogonal is as follows.

Definition 19.7.1 *Let U be a real $n \times n$ matrix. Then U is called orthogonal if $U^T U = U U^T = I$.*

Then the following proposition relates this to preservation of lengths of vectors.

Proposition 19.7.2 *An $n \times n$ matrix U is orthogonal if and only if $|Ux| = |x|$ for all vectors x .*

Proof: First suppose the matrix U preserves all lengths. Since U preserves distances, $|Uu| = |u|$ for every u . Let u, v be arbitrary vectors in \mathbb{R}^n and let $\theta \in \mathbb{R}$, $|\theta| = 1$, and $\theta(U^T U u - u, v) = |(U^T U u - u, v)|$. Therefore from the axioms of the inner product and Proposition 19.7.2,

$$\begin{aligned} |u|^2 + |v|^2 + 2\theta(u, v) &= |\theta u|^2 + |v|^2 + \theta(u, v) + \theta(v, u) \\ &= |\theta u + v|^2 = (U(\theta u + v), U(\theta u + v)) \\ &= (U\theta u, U\theta u) + (Uv, Uv) + (U\theta u, Uv) + (Uv, U\theta u) \\ &= |\theta u|^2 + |v|^2 + \theta(U^T U u, v) + \theta(v, U^T U u) \\ &= |u|^2 + |v|^2 + 2\theta(U^T U u, v) \end{aligned}$$

and so, subtracting the ends, it follows that for all u, v ,

$$0 = 2\theta(U^T U u - u, v) = 2|(U^T U u - u, v)|$$

from the above choice of θ . Now let $v = U^T U u - u$. It follows that

$$U^T U u - u = (U^T U - I)u = 0.$$

This is true for all u and so $U^T U = I$. Thus it is also true that $U U^T = I$.

Conversely, if $U^T U = I$, then

$$|Uu|^2 = (Uu, Uu) = (U^T U u, u) = (u, u) = |u|^2$$

Thus U preserves distance. ■

19.8 Diagonalization of Symmetric Matrices

Recall that a symmetric matrix is a real $n \times n$ matrix A such that $A^T = A$. One nice thing about symmetric matrices is that they have only real eigenvalues. You might want to review the property of the conjugate which says that $\overline{\bar{z}w} = z\bar{w}$ and how the conjugate of a sum is the sum of the conjugates.

Proposition 19.8.1 *Suppose A is a real symmetric matrix. Then all eigenvalues are real.*

Proof: Suppose $Ax = \lambda x$. Then

$$\bar{x}^T Ax = \bar{x}^T \lambda x = \lambda \bar{x}^T x = \lambda x^T \bar{x}$$

The last step happens because both $\bar{x}^T x$ and $x^T \bar{x}$ are the sum of the entries of x times the conjugate of these entries. Also $\bar{x}^T Ax$ is some complex number, a 1×1 matrix and so it equals its transpose. Thus, since $A = A^T$,

$$\bar{x}^T Ax = x^T A^T \bar{x} = x^T A \bar{x} = x^T \overline{A \bar{x}} = x^T \overline{\lambda x} = \bar{\lambda} x^T \bar{x}$$

Since $x \neq 0$, $x^T \bar{x}$ is a positive real number. Hence, the above shows that $\lambda = \bar{\lambda}$. ■

Definition 19.8.2 *A set of vectors in \mathbb{R}^p $\{x_1, \dots, x_k\}$ is called an **orthonormal set** of vectors if*

$$x_i^T x_j = \delta_{ij} \equiv \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

Note this is the same as saying that $(x_i, x_j) = x_i \cdot x_j = \delta_{ij}$.

What does it mean to say that $U^T U = I$ which is the definition for U to be orthogonal?

This says that for $U = \begin{pmatrix} u_1 & \cdots & u_n \end{pmatrix}$, $U^T = \begin{pmatrix} u_1^T \\ \vdots \\ u_n^T \end{pmatrix}$ and so from the way we multiply

matrices in which the ij^{th} entry of the product is the product of the i^{th} row of the matrix on the left with the j^{th} column of the matrix on the right, we have $u_i^T u_j = \delta_{ij}$. In other words, the columns of U are orthonormal. From this simple observation, the following theorem is obtained.

Theorem 19.8.3 *Let $\{u_1, \dots, u_n\}$ be orthonormal. Then it is linearly independent.*

Proof: We know from the above discussion that $U = \begin{pmatrix} u_1 & \cdots & u_n \end{pmatrix}$ is orthogonal. Thus if $Ux = 0$, you can multiply on the left on both sides with U^T and obtain $x = U^T Ux = U^T 0 = 0$. Thus, from the definition of linear independence, Definition 18.5.1, it follows that the columns of U comprise an independent set of vectors. ■

The proof of the following theorem is based on the Gram Schmidt process.

Theorem 19.8.4 *Let $\{x_1, \dots, x_n\}$ be linearly independent in \mathbb{R}^p , $p \geq n$. Then there exist orthonormal vectors $\{u_1, \dots, u_n\}$ which have the property that for each $k \leq n$, $\text{span}(x_1, \dots, x_k) = \text{span}(u_1, \dots, u_k)$.*

Proof: Let $\mathbf{u}_1 \equiv \mathbf{x}_1/|\mathbf{x}_1|$. Thus for $k = 1$, $\text{span}(\mathbf{u}_1) = \text{span}(\mathbf{x}_1)$ and $\{\mathbf{u}_1\}$ is an orthonormal set. Now suppose for some $k < n$, $\mathbf{u}_1, \dots, \mathbf{u}_k$ have been chosen such that $(\mathbf{u}_j, \mathbf{u}_l) = \delta_{jl}$ and $\text{span}(\mathbf{x}_1, \dots, \mathbf{x}_k) = \text{span}(\mathbf{u}_1, \dots, \mathbf{u}_k)$. Then define

$$\mathbf{u}_{k+1} \equiv \frac{\mathbf{x}_{k+1} - \sum_{j=1}^k (\mathbf{x}_{k+1} \cdot \mathbf{u}_j) \mathbf{u}_j}{\left| \mathbf{x}_{k+1} - \sum_{j=1}^k (\mathbf{x}_{k+1} \cdot \mathbf{u}_j) \mathbf{u}_j \right|}, \quad (19.2)$$

where the denominator is non-zero because the sum is in the span of the $\{\mathbf{x}_1, \dots, \mathbf{x}_k\}$. Thus by induction,

$$\mathbf{u}_{k+1} \in \text{span}(\mathbf{u}_1, \dots, \mathbf{u}_k, \mathbf{x}_{k+1}) = \text{span}(\mathbf{x}_1, \dots, \mathbf{x}_k, \mathbf{x}_{k+1}).$$

Also, $\mathbf{x}_{k+1} \in \text{span}(\mathbf{u}_1, \dots, \mathbf{u}_k, \mathbf{u}_{k+1})$ from solving 19.2 for \mathbf{x}_{k+1} , and it follows

$$\text{span}(\mathbf{x}_1, \dots, \mathbf{x}_k, \mathbf{x}_{k+1}) = \text{span}(\mathbf{u}_1, \dots, \mathbf{u}_k, \mathbf{u}_{k+1}).$$

If $l \leq k$,

$$\begin{aligned} (\mathbf{u}_{k+1} \cdot \mathbf{u}_l) &= C \left((\mathbf{x}_{k+1} \cdot \mathbf{u}_l) - \sum_{j=1}^k (\mathbf{x}_{k+1} \cdot \mathbf{u}_j) (\mathbf{u}_j \cdot \mathbf{u}_l) \right) = \\ &= C \left((\mathbf{x}_{k+1} \cdot \mathbf{u}_l) - \sum_{j=1}^k (\mathbf{x}_{k+1} \cdot \mathbf{u}_j) \delta_{lj} \right) = C((\mathbf{x}_{k+1} \cdot \mathbf{u}_l) - (\mathbf{x}_{k+1} \cdot \mathbf{u}_l)) = 0. \end{aligned}$$

The vectors, $\{\mathbf{u}_j\}_{j=1}^n$, generated in this way are therefore orthonormal because each vector has unit length. ■

Theorem 19.8.5 Let \mathbf{v}_1 be a unit vector ($|\mathbf{v}_1| = 1$) in \mathbb{R}^p , $p > 1$. Then there exist vectors

$$\{\mathbf{v}_2, \dots, \mathbf{v}_p\}$$

such that $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p\}$ is an orthonormal set of vectors.

Proof: Use Theorem 18.5.10 to extend $\{\mathbf{v}_1\}$ to a basis for \mathbb{R}^n and then use Theorem 19.8.4. ■

Thus, as observed above, the matrix $(\mathbf{v}_1 \ \dots \ \mathbf{v}_p)$ is a orthogonal matrix. With this preparation, here is a major result. It is actually a specialization of a much more interesting theorem. See any of my linear algebra books under the topic of Schur's theorem.

Theorem 19.8.6 Let A be a real symmetric matrix. Then there is an orthogonal transformation U such that

$$U^T A U = D$$

where D is a diagonal matrix having the real eigenvalues of A down the diagonal. Also, the columns of U are an orthonormal set of eigenvectors.

Proof: This is obviously true if A is a 1×1 matrix. Indeed, you let $U = 1$ and it all works because in this case A is already a diagonal matrix. Suppose then that the theorem is true for any $k < p$ and let A be a real $p \times p$ symmetric matrix. Then by the fundamental theorem of algebra, there exists a solution λ to the characteristic equation

$$\det(A - \lambda I) = 0.$$

Then since $A - \lambda I$ has no inverse, it follows that the columns are dependent and so there exists a nonzero vector \mathbf{u} such that $(A - \lambda I)\mathbf{u} = \mathbf{0}$ and from Proposition 19.8.1, λ is real. Dividing this vector by its magnitude, we can assume that $|\mathbf{u}| = 1$. By Theorem 19.8.5, there are vectors $\mathbf{v}_2, \dots, \mathbf{v}_p$ such that $\{\mathbf{u}, \mathbf{v}_2, \dots, \mathbf{v}_p\}$ is an orthonormal set of vectors. As observed above, if

$$U = (\mathbf{u} \quad \mathbf{v}_2 \quad \cdots \quad \mathbf{v}_p)$$

it follows that U is an orthogonal matrix. Now consider $U^T A U$. From the way we multiply matrices, this is

$$\begin{aligned} \begin{pmatrix} \mathbf{u}^T \\ \mathbf{v}_2^T \\ \vdots \\ \mathbf{v}_p^T \end{pmatrix} A (\mathbf{u} \quad \mathbf{v}_2 \quad \cdots \quad \mathbf{v}_p) &= \begin{pmatrix} \mathbf{u}^T \\ \mathbf{v}_2^T \\ \vdots \\ \mathbf{v}_p^T \end{pmatrix} (\mathbf{A}\mathbf{u} \quad \mathbf{A}\mathbf{v}_2 \quad \cdots \quad \mathbf{A}\mathbf{v}_p) \\ &= \begin{pmatrix} \mathbf{u}^T \\ \mathbf{v}_2^T \\ \vdots \\ \mathbf{v}_p^T \end{pmatrix} (\mathbf{A}\mathbf{u} \quad \mathbf{A}\mathbf{v}_2 \quad \cdots \quad \mathbf{A}\mathbf{v}_p) = \begin{pmatrix} \mathbf{u}^T \\ \mathbf{v}_2^T \\ \vdots \\ \mathbf{v}_p^T \end{pmatrix} (\lambda \mathbf{u} \quad \mathbf{A}\mathbf{v}_2 \quad \cdots \quad \mathbf{A}\mathbf{v}_p) \end{aligned}$$

Now recall the way we multiply matrices in which the ij^{th} entry is the product of the i^{th} row on the left with the j^{th} column on the right. Thus, since these columns of U are orthonormal, the above product reduces to something of the form

$$\begin{pmatrix} \lambda & \mathbf{a}^T \\ \mathbf{0} & A_1 \end{pmatrix}$$

where A_1 is an $(p-1) \times (p-1)$ matrix. Summarizing, there is an orthogonal matrix U such that

$$U^T A U = \begin{pmatrix} \lambda & \mathbf{a}^T \\ \mathbf{0} & A_1 \end{pmatrix}$$

I claim that $\mathbf{a} = \mathbf{0}$. To see this, take the transpose of both sides, using symmetry of A to obtain

$$\begin{pmatrix} \lambda & \mathbf{a}^T \\ \mathbf{0} & A_1 \end{pmatrix} = U^T A U = (U^T A U)^T = \begin{pmatrix} \lambda & \mathbf{0}^T \\ \mathbf{a} & A_1 \end{pmatrix}$$

Thus $\mathbf{a} = \mathbf{0}$ as claimed. Now by induction, there is an orthogonal matrix \hat{U} such that

$$\hat{U}^T A_1 \hat{U} = D$$

where D is a diagonal matrix. Now note that from the way we multiply matrices,

$$\begin{aligned} \begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \hat{U} \end{pmatrix}^T \begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \hat{U} \end{pmatrix} &= \begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \hat{U}^T \end{pmatrix} \begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \hat{U} \end{pmatrix} \\ &= \begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \hat{U}^T \end{pmatrix} \begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \hat{U} \end{pmatrix} = \begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \hat{U}^T \hat{U} \end{pmatrix} = \begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & I \end{pmatrix} = I \end{aligned}$$

Thus

$$\begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \hat{U} \end{pmatrix} \equiv \tilde{U}$$

is an orthogonal matrix. Now

$$\begin{aligned}\tilde{U}^T U^T A U \tilde{U} &= \tilde{U}^T \begin{pmatrix} \lambda & \mathbf{0}^T \\ \mathbf{0} & A_1 \end{pmatrix} \tilde{U} \\ &= \begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \hat{U}^T \end{pmatrix} \begin{pmatrix} \lambda & \mathbf{0}^T \\ \mathbf{0} & A_1 \end{pmatrix} \begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \hat{U} \end{pmatrix} \\ &= \begin{pmatrix} \lambda & \mathbf{0}^T \\ \mathbf{0} & \hat{U}^T A_1 \hat{U} \end{pmatrix} = \begin{pmatrix} \lambda & \mathbf{0}^T \\ \mathbf{0} & D \end{pmatrix}\end{aligned}$$

which is a diagonal matrix. This shows the first part. Now if

$$U = (\mathbf{u}_1 \quad \mathbf{u}_2 \quad \cdots \quad \mathbf{u}_p)$$

$$D = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix}, U^T A U = D$$

then, multiplying on both sides by U ,

$$AU = UD$$

and so, from the way we multiply matrices, this yields

$$\begin{aligned}AU &= (\mathbf{A}\mathbf{u}_1 \quad \mathbf{A}\mathbf{u}_2 \quad \cdots \quad \mathbf{A}\mathbf{u}_p) = UD \\ &= (\lambda_1 \mathbf{u}_1 \quad \lambda_2 \mathbf{u}_2 \quad \cdots \quad \lambda_p \mathbf{u}_p)\end{aligned}$$

which shows that $\mathbf{A}\mathbf{u}_j = \lambda_j \mathbf{u}_j$ for each j . This shows the columns of U form an orthonormal set of eigenvectors and the diagonal entries of D are the eigenvalues of A . ■

Example 19.8.7 Here is a symmetric matrix which has eigenvalues 6, -12 , 18

$$A = \begin{pmatrix} 1 & -4 & 13 \\ -4 & 10 & -4 \\ 13 & -4 & 1 \end{pmatrix}$$

Find a matrix U such that $U^T A U$ is a diagonal matrix.

From the above explanation the columns of this matrix U are eigenvectors of unit length and in fact this is sufficient to obtain the matrix. After doing row operations to find the eigenvectors and then dividing each by its magnitude, you obtain

$$\begin{pmatrix} 1 & -4 & 13 \\ -4 & 10 & -4 \\ 13 & -4 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{6}\sqrt{6} \\ \frac{1}{3}\sqrt{6} \\ \frac{1}{6}\sqrt{6} \end{pmatrix} = \begin{pmatrix} \sqrt{6} \\ 2\sqrt{6} \\ \sqrt{6} \end{pmatrix} = 6 \begin{pmatrix} \frac{1}{6}\sqrt{6} \\ \frac{1}{3}\sqrt{6} \\ \frac{1}{6}\sqrt{6} \end{pmatrix}$$

$$\begin{pmatrix} 1 & -4 & 13 \\ -4 & 10 & -4 \\ 13 & -4 & 1 \end{pmatrix} \begin{pmatrix} -\frac{1}{2}\sqrt{2} \\ 0 \\ \frac{1}{2}\sqrt{2} \end{pmatrix} = \begin{pmatrix} 6\sqrt{2} \\ 0 \\ -6\sqrt{2} \end{pmatrix} = -12 \begin{pmatrix} -\frac{1}{2}\sqrt{2} \\ 0 \\ \frac{1}{2}\sqrt{2} \end{pmatrix}$$

$$\begin{pmatrix} 1 & -4 & 13 \\ -4 & 10 & -4 \\ 13 & -4 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{3}\sqrt{3} \\ -\frac{1}{3}\sqrt{3} \\ \frac{1}{3}\sqrt{3} \end{pmatrix} = \begin{pmatrix} 6\sqrt{3} \\ -6\sqrt{3} \\ 6\sqrt{3} \end{pmatrix} = 18 \begin{pmatrix} \frac{1}{3}\sqrt{3} \\ -\frac{1}{3}\sqrt{3} \\ \frac{1}{3}\sqrt{3} \end{pmatrix}$$

Thus the matrix of interest is

$$U = \begin{pmatrix} \frac{1}{6}\sqrt{6} & -\frac{1}{2}\sqrt{2} & \frac{1}{3}\sqrt{3} \\ \frac{1}{3}\sqrt{6} & 0 & -\frac{1}{3}\sqrt{3} \\ \frac{1}{6}\sqrt{6} & \frac{1}{2}\sqrt{2} & \frac{1}{3}\sqrt{3} \end{pmatrix}$$

Then

$$\begin{aligned} & \begin{pmatrix} \frac{1}{6}\sqrt{6} & -\frac{1}{2}\sqrt{2} & \frac{1}{3}\sqrt{3} \\ \frac{1}{3}\sqrt{6} & 0 & -\frac{1}{3}\sqrt{3} \\ \frac{1}{6}\sqrt{6} & \frac{1}{2}\sqrt{2} & \frac{1}{3}\sqrt{3} \end{pmatrix}^T \begin{pmatrix} 1 & -4 & 13 \\ -4 & 10 & -4 \\ 13 & -4 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{6}\sqrt{6} & -\frac{1}{2}\sqrt{2} & \frac{1}{3}\sqrt{3} \\ \frac{1}{3}\sqrt{6} & 0 & -\frac{1}{3}\sqrt{3} \\ \frac{1}{6}\sqrt{6} & \frac{1}{2}\sqrt{2} & \frac{1}{3}\sqrt{3} \end{pmatrix} \\ &= \begin{pmatrix} 6 & 0 & 0 \\ 0 & -12 & 0 \\ 0 & 0 & 18 \end{pmatrix} \end{aligned}$$

19.9 Exercises

1. Let

$$\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$$

be a basis for \mathbb{F}^n and define a mapping $T : \mathbb{F}^n \rightarrow \text{span}(\mathbf{v}_1, \dots, \mathbf{v}_r)$ as follows.

$$T\left(\sum_{k=1}^n a_k \mathbf{u}_k\right) \equiv \sum_{k=1}^r a_k \mathbf{v}_k$$

Explain why this is a linear transformation.

2. In the above problem, suppose $\mathbf{v}_k = \mathbf{u}_k$. Show $T\mathbf{v} = \mathbf{v}$ if $\mathbf{v} \in V \equiv \text{span}(\mathbf{u}_1, \dots, \mathbf{u}_r)$. Now show that $T(T(\mathbf{x})) = T(\mathbf{x})$.
3. The Cayley Hamilton theorem states that every matrix satisfies its characteristic equation. I have given a short proof of this major theorem in the appendix on the theory of determinants. See Section 20.2.10. Suppose you have $p(\lambda)$ is the characteristic polynomial for a square $n \times n$ matrix A . Show that this matrix is invertible if and only if the constant term of the $p(\lambda)$ is non zero. In this case, give a formula for A^{-1} in terms of powers of A . Say

$$p(\lambda) = \lambda^n + a_{n-1}\lambda^{n-1} + \dots + a_1\lambda + a_0$$

Thus you need explain why $a_0 \neq 0$ if A^{-1} exists and then find a formula for A^{-1} when this is the case. **Hint:** By the Cayley Hamilton theorem $p(A) = 0$ meaning

$$A^n + a_{n-1}A^{n-1} + \dots + a_1A + a_0I = 0$$

Now consider solving for I and factoring out A .

4. Here are some matrices. Label according to whether they are symmetric, skew symmetric, or orthogonal. If the matrix is orthogonal, determine whether it is proper or improper.

$$(a) \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/\sqrt{2} & -1/\sqrt{2} \\ 0 & 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \quad (b) \begin{pmatrix} 1 & 2 & -3 \\ 2 & 1 & 4 \\ -3 & 4 & 7 \end{pmatrix} \quad (c) \begin{pmatrix} 0 & -2 & -3 \\ 2 & 0 & -4 \\ 3 & 4 & 0 \end{pmatrix}$$

5. Show that every real matrix may be written as the sum of a skew symmetric and a symmetric matrix. **Hint:** If A is an $n \times n$ matrix, show that $B \equiv \frac{1}{2}(A - A^T)$ is skew symmetric.
6. Let \mathbf{x} be a vector in \mathbb{R}^n and consider the matrix $I - \frac{2\mathbf{x}\mathbf{x}^T}{|\mathbf{x}|^2}$. Show this matrix is both symmetric and orthogonal.
7. For U an orthogonal matrix, explain why $|U\mathbf{x}| = |\mathbf{x}|$ for any vector \mathbf{x} . Next explain why if U is an $n \times n$ matrix with the property that $|U\mathbf{x}| = |\mathbf{x}|$ for all vectors, \mathbf{x} , then U must be orthogonal. Thus the orthogonal matrices are exactly those which preserve distance. This was done in general in the chapter for orthogonal matrices. Try to do it in your own words.
8. A quadratic form in three variables is an expression of the form $a_1x^2 + a_2y^2 + a_3z^2 + a_4xy + a_5xz + a_6yz$. Show that every such quadratic form may be written as

$$\begin{pmatrix} x & y & z \end{pmatrix} A \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

where A is a symmetric matrix.

9. Given a quadratic form in three variables, x, y , and z , show there exists an orthogonal matrix U and variables x', y', z' such that $\begin{pmatrix} x & y & z \end{pmatrix}^T = U \begin{pmatrix} x' & y' & z' \end{pmatrix}^T$ with the property that in terms of the new variables, the quadratic form is

$$\lambda_1 (x')^2 + \lambda_2 (y')^2 + \lambda_3 (z')^2$$

where the numbers, λ_1, λ_2 , and λ_3 are the eigenvalues of the matrix A in Problem 8.

10. If A is a symmetric invertible matrix, is it always the case that A^{-1} must be symmetric also? How about A^k for k a positive integer? Explain.
11. If A, B are symmetric matrices, does it follow that AB is also symmetric?
12. Suppose A, B are symmetric and $AB = BA$. Does it follow that AB is symmetric?
13. Here are some matrices. What can you say about the eigenvalues of these matrices just by looking at them?

$$(a) \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}$$

$$(c) \begin{pmatrix} 0 & -2 & -3 \\ 2 & 0 & -4 \\ 3 & 4 & 0 \end{pmatrix}$$

$$(b) \begin{pmatrix} 1 & 2 & -3 \\ 2 & 1 & 4 \\ -3 & 4 & 7 \end{pmatrix}$$

$$(d) \begin{pmatrix} 1 & 2 & 3 \\ 0 & 2 & 3 \\ 0 & 0 & 2 \end{pmatrix}$$

14. Find the eigenvalues and eigenvectors of the matrix $\begin{pmatrix} c & 0 & 0 \\ 0 & 0 & -b \\ 0 & b & 0 \end{pmatrix}$. Here b, c are real numbers.

15. Find the eigenvalues and eigenvectors of the matrix $\begin{pmatrix} c & 0 & 0 \\ 0 & a & -b \\ 0 & b & a \end{pmatrix}$. Here a, b, c are real numbers.

16. Find the eigenvalues and an orthonormal basis of eigenvectors for A .

$$A = \begin{pmatrix} 11 & -1 & -4 \\ -1 & 11 & -4 \\ -4 & -4 & 14 \end{pmatrix}.$$

Hint: Two eigenvalues are 12 and 18.

17. Find the eigenvalues and an orthonormal basis of eigenvectors for A .

$$A = \begin{pmatrix} 4 & 1 & -2 \\ 1 & 4 & -2 \\ -2 & -2 & 7 \end{pmatrix}.$$

Hint: One eigenvalue is 3.

18. Show that if A is a real symmetric matrix and λ and μ are two different eigenvalues, then if \mathbf{x} is an eigenvector for λ and \mathbf{y} is an eigenvector for μ , then $\mathbf{x} \cdot \mathbf{y} = 0$. Also all eigenvalues are real. Supply reasons for each step in the following argument. First

$$\lambda \mathbf{x}^T \bar{\mathbf{x}} = (A\mathbf{x})^T \bar{\mathbf{x}} = \mathbf{x}^T A \bar{\mathbf{x}} = \mathbf{x}^T \overline{A\mathbf{x}} = \mathbf{x}^T \overline{\lambda \mathbf{x}} = \overline{\lambda} \mathbf{x}^T \bar{\mathbf{x}}$$

and so $\lambda = \overline{\lambda}$. This shows that all eigenvalues are real. It follows all the eigenvectors are real. Why? Now let $\mathbf{x}, \mathbf{y}, \mu$ and λ be given as above.

$$\lambda (\mathbf{x} \cdot \mathbf{y}) = \lambda \mathbf{x} \cdot \mathbf{y} = A\mathbf{x} \cdot \mathbf{y} = \mathbf{x} \cdot A\mathbf{y} = \mathbf{x} \cdot \mu \mathbf{y} = \mu (\mathbf{x} \cdot \mathbf{y})$$

and so

$$(\lambda - \mu) (\mathbf{x} \cdot \mathbf{y}) = 0.$$

Since $\lambda \neq \mu$, it follows $\mathbf{x} \cdot \mathbf{y} = 0$.

19. Suppose U is an orthogonal $n \times n$ matrix. Explain why $\text{rank}(U) = n$.

20. Show that the eigenvalues and eigenvectors of a real matrix occur in conjugate pairs.

21. If a real matrix A has all real eigenvalues, does it follow that A must be symmetric. If so, explain why and if not, give an example to the contrary.
22. Suppose A is a 3×3 symmetric matrix and you have found two eigenvectors which form an orthonormal set. Explain why their cross product is also an eigenvector.
23. Determine which of the following sets of vectors are orthonormal sets. Justify your answer.

(a) $\{(1, 1), (1, -1)\}$

(b) $\left\{\left(\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}\right), (1, 0)\right\}$

(c) $\left\{\left(\frac{1}{3}, \frac{2}{3}, \frac{2}{3}\right), \left(\frac{-2}{3}, \frac{-1}{3}, \frac{2}{3}\right), \left(\frac{2}{3}, \frac{-2}{3}, \frac{1}{3}\right)\right\}$

24. Show that if $\{u_1, \dots, u_n\}$ is an orthonormal set of vectors in \mathbb{R}^n , then it is a basis.

Hint: It was shown earlier that this is a linearly independent set.

25. Fill in the missing entries to make the matrix orthogonal.

$$\begin{pmatrix} \frac{-1}{\sqrt{2}} & \frac{-1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & - & - \\ - & \frac{\sqrt{6}}{3} & - \end{pmatrix}.$$

26. Fill in the missing entries to make the matrix orthogonal.

$$\begin{pmatrix} \frac{1}{3} & -\frac{2}{\sqrt{5}} & - \\ \frac{2}{3} & 0 & - \\ - & - & \frac{4}{15}\sqrt{5} \end{pmatrix}$$

27. Find the eigenvalues and an orthonormal basis of eigenvectors for A . Diagonalize A by finding an orthogonal matrix U and a diagonal matrix D such that $U^T A U = D$.

$$A = \begin{pmatrix} -1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{pmatrix}.$$

Hint: One eigenvalue is -2.

28. Find the eigenvalues and an orthonormal basis of eigenvectors for A . Diagonalize A by finding an orthogonal matrix U and a diagonal matrix D such that $U^T A U = D$.

$$A = \begin{pmatrix} 17 & -7 & -4 \\ -7 & 17 & -4 \\ -4 & -4 & 14 \end{pmatrix}.$$

Hint: Two eigenvalues are 18 and 24.

29. Find the eigenvalues and an orthonormal basis of eigenvectors for A . Diagonalize A by finding an orthogonal matrix U and a diagonal matrix D such that $U^T A U = D$.

$$A = \begin{pmatrix} 13 & 1 & 4 \\ 1 & 13 & 4 \\ 4 & 4 & 10 \end{pmatrix}.$$

Hint: Two eigenvalues are 12 and 18.

30. Find the eigenvalues and an orthonormal basis of eigenvectors for A . Diagonalize A by finding an orthogonal matrix U and a diagonal matrix D such that $U^T A U = D$.

$$A = \begin{pmatrix} 3 & 0 & 0 \\ 0 & \frac{3}{2} & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{3}{2} \end{pmatrix}.$$

31. Find the eigenvalues and an orthonormal basis of eigenvectors for A . Diagonalize A by finding an orthogonal matrix U and a diagonal matrix D such that $U^T A U = D$.

$$A = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 5 & 1 \\ 0 & 1 & 5 \end{pmatrix}.$$

32. Explain why a real matrix A is symmetric if and only if there exists an orthogonal matrix U such that $A = U^T D U$ for D a diagonal matrix.

33. Find an orthonormal basis for the spans of the following sets of vectors.

(a) $(3, -4, 0), (7, -1, 0), (1, 7, 1)$.

(b) $(3, 0, -4), (11, 0, 2), (1, 1, 7)$

(c) $(3, 0, -4), (5, 0, 10), (-7, 1, 1)$

34. The set, $V \equiv \{(x, y, z) : 2x + 3y - z = 0\}$ is a subspace of \mathbb{R}^3 . Find an orthonormal basis for this subspace.

35. The two level surfaces, $2x + 3y - z + w = 0$ and $3x - y + z + 2w = 0$ intersect in a subspace of \mathbb{R}^4 , find a basis for this subspace. Next find an orthonormal basis for this subspace.

36. Let A, B be a $m \times n$ matrices. Define an inner product on the set of real $m \times n$ matrices by

$$(A, B)_F \equiv \text{trace}(AB^T).$$

Show this is an inner product satisfying all the inner product axioms. Recall for M an $n \times n$ matrix, $\text{trace}(M) \equiv \sum_{i=1}^n M_{ii}$. The resulting norm, $\|\cdot\|_F$ is called the Frobenius norm and it can be used to measure the distance between two matrices.

37. The trace of an $n \times n$ matrix M is defined as $\sum_i M_{ii}$. In other words it is the sum of the entries on the main diagonal. If A, B are $n \times n$ matrices, show $\text{trace}(AB) = \text{trace}(BA)$. Now explain why if $A = S^{-1}BS$ it follows $\text{trace}(A) = \text{trace}(B)$. **Hint:** For the first part, write these in terms of components of the matrices and it just falls out.

38. For U a matrix, a number will be called $o(U)$ if it satisfies $\lim_{\|U\| \rightarrow 0} \frac{o(U)}{\|U\|} = 0$. Here $\|U\|$ will be the Frobenius norm of U . Show that for U an $n \times n$ matrix, $\det(I + U) = 1 + \text{trace}(U) + o(U)$. Explain why if a number is a product of more than one entry of U then it must be $o(U)$. For example, $U_{12}U_{23}$ would be $o(U)$. **Hint:** This is true obviously if $n = 1$. Suppose true for $n - 1$ and expand along last column and use induction to get the result for n .

39. Next show that if F^{-1} exists, then

$$\det(F + U) - \det(F) = \det(F) \text{trace}(F^{-1}U) + o(U)$$

Hint: Factor out F from $F + U$.

40. Let $A(t)$ be an $m \times n$ matrix whose entries are differentiable functions of t . The symbol $A'(t)$, means to replace each t dependent entry of $A(t)$ with its derivative. Thus if

$$A(t) = \begin{pmatrix} \sin t & t^2 \\ t+1 & \ln(1+t^2) \end{pmatrix}, \text{ then } A'(t) = \begin{pmatrix} \cos t & 2t \\ 1 & 2\frac{t}{t^2+1} \end{pmatrix}$$

Let $A(t)$ be an $m \times n$ matrix and let $B(t)$ be an $n \times p$ matrix. Show the product rule.

$$(AB)'(t) = A'(t)B(t) + A(t)B'(t)$$

Hint: Just use the entries of both sides and reduce to the usual product rule. That is, the ij^{th} entry of $(AB)'(t)$ is $\sum_k (A_{ik}B_{kj})'(t)$. Now use the product rule.

Chapter 20

The Mathematical Theory of Determinants*



You might skip this chapter and return to it later if accepting the outrageous claims about the determinant, that it is independent of the row or column chosen, does not bother you. If this does cause some cognitive dissonance, then you should read this chapter now.

20.1 The Function sgn

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

Lemma 20.1.1 *There exists a function, sgn_n which maps each ordered list of numbers from $\{1, \dots, n\}$ to one of the three numbers, 0, 1, or -1 which also has the following properties.*

$$\text{sgn}_n(1, \dots, n) = 1 \quad (20.1)$$

$$\text{sgn}_n(i_1, \dots, p, \dots, q, \dots, i_n) = -\text{sgn}_n(i_1, \dots, q, \dots, p, \dots, i_n) \quad (20.2)$$

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 $\{i_1, \dots, i_n\} = \{1, \dots, n\}$ so that every number from $\{1, \dots, n\}$ appears in the ordered list, (i_1, \dots, i_n) ,

$$\begin{aligned} \text{sgn}_n(i_1, \dots, i_{\theta-1}, n, i_{\theta+1}, \dots, i_n) &\equiv \\ (-1)^{n-\theta} \text{sgn}_{n-1}(i_1, \dots, i_{\theta-1}, i_{\theta+1}, \dots, i_n) \end{aligned} \quad (20.3)$$

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

Proof: Define $\text{sign}(x) = 1$ 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 $\text{sgn}_1(1) \equiv 1$. For the general case where $n > 1$, simply define

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

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, \dots, i_n) are switched. Denote the new ordered list of numbers as (j_1, \dots, 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

$$\begin{array}{ccccccc} \frac{i_1}{1} & \frac{i_2}{2} & \dots & \frac{i_p}{p} & \dots & \frac{i_q}{q} & \dots & \frac{i_n}{n} \\ \\ \frac{i_1}{1} & \frac{i_2}{2} & \dots & \frac{i_q}{p} & \dots & \frac{i_p}{q} & \dots & \frac{i_n}{n} \\ \\ \frac{j_1}{1} & \frac{j_2}{2} & \dots & \frac{j_p}{p} & \dots & \frac{j_q}{q} & \dots & \frac{j_n}{n} \end{array}$$

Then

$$\begin{aligned} \operatorname{sgn}_n(j_1, \dots, j_n) &\equiv \operatorname{sign} \left(\prod_{r < s} (j_s - j_r) \right) \\ &= \operatorname{sign} \left(\overbrace{(i_p - i_q) \prod_{p < j < q} (i_j - i_q) \prod_{p < j < q} (i_p - i_j)}^{\text{one of } p, q} \prod_{r < s, r, s \notin \{p, q\}} (i_s - i_r) \right) \end{aligned}$$

The last product consists of the product of terms which were in the un-switched product $\prod_{r < s} (i_s - i_r)$ so produces no change in sign, while the two products in the middle both introduce $q - p - 1$ minus signs. Thus their product produces no change in sign. The first factor is of opposite sign to the $i_q - i_p$ which occurred in $\operatorname{sgn}_n(i_1, \dots, i_n)$. Therefore, this switch introduced a minus sign and

$$\operatorname{sgn}_n(j_1, \dots, j_n) = -\operatorname{sgn}_n(i_1, \dots, i_n)$$

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

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

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

$$\operatorname{sgn}_n(i_1, \dots, i_{\theta-1}, n, i_{\theta+1}, \dots, i_n) = (-1)^{n-\theta} \operatorname{sgn}_{n-1}(i_1, \dots, i_{\theta-1}, i_{\theta+1}, \dots, i_n)$$

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

Lemma 20.1.2 *Every ordered list of distinct numbers from $\{1, 2, \dots, n\}$ can be obtained from every other such ordered list 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 20.1 and 20.2, you could start with $f(1, \dots, n) = g(1, \dots, n) = 1$ and applying the same sequence of switches, eventually arrive at $f(i_1, \dots, i_n) = g(i_1, \dots, i_n)$. If any numbers are repeated, then 20.2 gives both functions are equal to zero for that ordered list. ■

Definition 20.1.3 When you have an ordered list of distinct numbers selected from $\{1, 2, \dots, n\}$, say (i_1, \dots, i_n) , this ordered list is called a permutation. The symbol for all such permutations is S_n . The number $\text{sgn}_n(i_1, \dots, i_n)$ is called the sign of the permutation.

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

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

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 .

20.2 The Determinant

Definition 20.2.1 Let f be a function which has the set of ordered lists of numbers from $\{1, \dots, n\}$ as its domain. Define

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

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

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

20.2.1 The Definition

Definition 20.2.2 Let $(a_{ij}) = A$ denote an $n \times n$ matrix. The determinant of A , denoted by $\det(A)$ is defined by

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

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

20.2.2 Permuting Rows Or Columns

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

$$\det(A(r_1, \dots, r_n)) = \sum_{(k_1, \dots, k_n)} \operatorname{sgn}(k_1, \dots, k_n) a_{r_1 k_1} \cdots a_{r_n k_n} \quad (20.4)$$

and

$$A(1, \dots, n) = A.$$

Proposition 20.2.3 *Let*

$$(r_1, \dots, r_n)$$

be an ordered list of numbers from $\{1, \dots, n\}$. Then

$$\operatorname{sgn}(r_1, \dots, r_n) \det(A)$$

$$= \sum_{(k_1, \dots, k_n)} \operatorname{sgn}(k_1, \dots, k_n) a_{r_1 k_1} \cdots a_{r_n k_n} \quad (20.5)$$

$$= \det(A(r_1, \dots, r_n)). \quad (20.6)$$

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

$$\det(A(1, \dots, r, \dots, s, \dots, n)) = \quad (20.7)$$

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

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

$$\begin{aligned} &= \sum_{(k_1, \dots, k_n)} \operatorname{sgn}(k_1, \dots, k_s, \dots, k_r, \dots, k_n) a_{1k_1} \cdots a_{rk_s} \cdots a_{sk_r} \cdots a_{nk_n} \\ &= \sum_{(k_1, \dots, k_n)} -\operatorname{sgn} \left(k_1, \dots, \overbrace{k_r, \dots, k_s}^{\text{These got switched}}, \dots, k_n \right) a_{1k_1} \cdots a_{sk_r} \cdots a_{rk_s} \cdots a_{nk_n} \\ &= -\det(A(1, \dots, s, \dots, r, \dots, n)). \end{aligned} \quad (20.8)$$

Consequently,

$$\begin{aligned} \det(A(1, \dots, s, \dots, r, \dots, n)) &= \\ -\det(A(1, \dots, r, \dots, s, \dots, n)) &= -\det(A) \end{aligned}$$

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

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

where it took p switches to obtain (r_1, \dots, r_n) from $(1, \dots, n)$. By Lemma 20.1.1, this implies

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

and proves the proposition in the case when there are no repeated numbers in the ordered list, (r_1, \dots, r_n) . However, if there is a repeat, say the r^{th} row equals the s^{th} row, then the reasoning of 20.7-20.8 shows that $\det A(r_1, \dots, r_n) = 0$ and also $\text{sgn}(r_1, \dots, r_n) = 0$ so the formula holds in this case also. ■

Observation 20.2.4 *There are $n!$ ordered lists of distinct numbers from $\{1, \dots, n\}$.*

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 $\{1, \dots, n\}$ as stated in the observation.

20.2.3 A Symmetric Definition

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

Corollary 20.2.5 *The following formula for $\det(A)$ is valid.*

$$\det(A) = \frac{1}{n!} \sum_{(r_1, \dots, r_n)} \sum_{(k_1, \dots, k_n)} \text{sgn}(r_1, \dots, r_n) \text{sgn}(k_1, \dots, k_n) a_{r_1 k_1} \cdots a_{r_n k_n}. \quad (20.9)$$

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

Proof: From Proposition 20.2.3, if the r_i are distinct,

$$\det(A) = \sum_{(k_1, \dots, k_n)} \text{sgn}(r_1, \dots, r_n) \text{sgn}(k_1, \dots, k_n) a_{r_1 k_1} \cdots a_{r_n k_n}.$$

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

$$n! \det(A) =$$

$$\sum_{(r_1, \dots, r_n)} \sum_{(k_1, \dots, k_n)} \text{sgn}(r_1, \dots, r_n) \text{sgn}(k_1, \dots, k_n) a_{r_1 k_1} \cdots a_{r_n k_n}.$$

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

20.2.4 The Alternating Property of the Determinant

Corollary 20.2.6 *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 $(xa_1 + yb_1, \dots, xa_n + yb_n)$. Then

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

where the i^{th} row of A_1 is (a_1, \dots, a_n) and the i^{th} row of A_2 is (b_1, \dots, 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 20.2.3 when two rows are switched, the determinant of the resulting matrix is (-1) times the determinant of the original matrix. By Corollary 20.2.5 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,

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

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.

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

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

20.2.5 Linear Combinations and Determinants

Linear combinations have been discussed already. However, here is a review and some new terminology.

Definition 20.2.7 A vector w , is a linear combination of the vectors

$$\{v_1, \dots, v_r\}$$

if there exists scalars, c_1, \dots, c_r such that $w = \sum_{k=1}^r c_k v_k$. This is the same as saying

$$w \in \operatorname{span}(v_1, \dots, v_r).$$

The following corollary is also of great use.

Corollary 20.2.8 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 $A = (a_1 \cdots a_n)$ be the columns of A and suppose the condition that one column is a linear combination of r of the others is satisfied. Then by using Corollary 20.2.6 the determinant of A is zero if and only if the determinant of the matrix B , which has this special column placed in the last position, equals zero. Thus $a_n = \sum_{k=1}^r c_k a_k$ and so

$$\det(B) = \det(a_1 \cdots a_r \cdots a_{n-1} \sum_{k=1}^r c_k a_k).$$

By Corollary 20.2.6

$$\det(B) = \sum_{k=1}^r c_k \det(a_1 \cdots a_r \cdots a_{n-1} a_k) = 0.$$

because there are two equal columns. The case for rows follows from the fact that $\det(A) = \det(A^T)$. ■

20.2.6 The Determinant of a Product

Recall the following definition of matrix multiplication.

Definition 20.2.9 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 20.2.10 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 20.2.3,

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

20.2.7 Cofactor Expansions

Lemma 20.2.11 Suppose a matrix is of the form

$$M = \begin{pmatrix} A & * \\ \mathbf{0} & a \end{pmatrix} \quad (20.10)$$

or

$$M = \begin{pmatrix} A & \mathbf{0} \\ * & a \end{pmatrix} \quad (20.11)$$

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,

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

Letting θ denote the position of n in the ordered list, (k_1, \dots, k_n) then using Lemma 20.1.1, $\det(M)$ equals

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

Now suppose 20.11. 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

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

To get the assertion in the situation of 20.10 use Corollary 20.2.5 and 20.11 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 20.2.12 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 20.2.13 *Let A be an $n \times n$ matrix where $n \geq 2$. Then*

$$\det(A) = \sum_{j=1}^n a_{ij} \operatorname{cof}(A)_{ij} = \sum_{i=1}^n a_{ij} \operatorname{cof}(A)_{ij}. \quad (20.12)$$

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}, \dots, 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

$$(0, \dots, 0, a_{ij}, 0, \dots, 0).$$

Then by Corollary 20.2.6,

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

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 $\operatorname{cof}(A)_{ij} \equiv (-1)^{i+j} \det(A^{ij})$. At this point, recall that from Proposition 20.2.3, 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 20.2.11,

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

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,

$$\begin{aligned} \det(A) &= \det(A^T) = \sum_{j=1}^n a_{ij}^T \operatorname{cof}(A^T)_{ij} \\ &= \sum_{j=1}^n a_{ji} \operatorname{cof}(A)_{ji} \end{aligned}$$

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

20.2.8 Row, Column, and Determinant Rank

This section will consider the concept of rank of a matrix. This is a number and its description is in the following definition.

Definition 20.2.14 *A sub-matrix 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 $r \times r$ sub-matrix 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 20.2.15 *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 $l \notin \{i_1, \dots, i_r\}$ 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 $p \notin \{j_1, \dots, j_r\}$ then the above would be a submatrix of A which is too large to have non zero determinant. On the other hand, if $p \in \{j_1, \dots, j_r\}$ 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 $a_{l p}$. This is given to be nonzero, it being the determinant of the matrix

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

Thus $0 = a_{l p} C + \sum_{k=1}^r C_k a_{i_k p}$ which implies $a_{l p} = \sum_{k=1}^r \frac{-C_k}{C} a_{i_k p} \equiv \sum_{k=1}^r m_k a_{i_k p}$. 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, \dots, i_r rows. ■

Corollary 20.2.16 *The determinant rank equals the row rank.*

Proof: From Theorem 20.2.15, the row rank is no larger than the determinant rank. Could the row rank be smaller than the determinant rank? If so, there exist p rows for $p < r$ such that the span of these p rows equals the row space. But this implies that the $r \times r$ submatrix whose determinant is nonzero also has row rank no larger than p which is impossible if its determinant is to be nonzero because at least one row is a linear combination of the others. ■

Corollary 20.2.17 *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 20.2.16, column rank of A = row rank of A^T = determinant rank of A^T = determinant rank of A . ■

20.2.9 Formula for the Inverse

Note that this gives an easy way to write a formula for the inverse of an $n \times n$ matrix.

Theorem 20.2.18 A^{-1} exists if and only if $\det(A) \neq 0$. If $\det(A) \neq 0$, then $A^{-1} = (a_{ij}^{-1})$ 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 20.2.13 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 consider

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

when $k \neq r$. Replace the k^{th} column with the r^{th} column to obtain a matrix B_k whose determinant equals zero by Corollary 20.2.6. However, expanding this matrix along the k^{th} column yields

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

Summarizing,

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

Using the other formula in Theorem 20.2.13, 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 $A^{-1} = (a_{ij}^{-1})$, where

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

Now suppose A^{-1} exists. Then by Theorem 20.2.10,

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

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 20.2.19 *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 20.2.10 implies

$$\det B \det A = 1$$

and so $\det A \neq 0$. Therefore from Theorem 20.2.18, A^{-1} exists. Therefore,

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

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 20.2.18 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 .

20.2.10 The Cayley Hamilton Theorem

Definition 20.2.20 *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 algebra¹. 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 20.2.21 *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 $A_i = 0$.

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 $\lambda \rightarrow \infty$. Thus $A_p = 0$ after all. Hence each $A_i = 0$. ■

With the lemma, here is a simple corollary.

¹A 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.

Corollary 20.2.22 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 20.2.23 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 20.2.18

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

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

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

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

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

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

$$\begin{aligned} a_0I + a_1I\lambda + \cdots + I\lambda^n &= C_0\lambda + C_1\lambda^2 + \cdots + C_{n-1}\lambda^n \\ &\quad - [C_0A + C_1A\lambda + \cdots + C_{n-1}A\lambda^{n-1}] \\ &= -C_0A + (C_0 - C_1A)\lambda + (C_1 - C_2A)\lambda^2 + \cdots + (C_{n-2} - C_{n-1}A)\lambda^{n-1} + C_{n-1}\lambda^n \end{aligned}$$

Then, using Corollary 20.2.22, 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. ■

20.2.11 Cramer's Rule

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

$$\mathbf{x} = (A^{-1}A)\mathbf{x} = A^{-1}(A\mathbf{x}) = A^{-1}\mathbf{y}$$

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 $(y_1, \dots, y_n)^T$, and the determinant of this modified matrix is taken and divided by $\det(A)$. This formula is known as Cramer's rule.

20.3 p Dimensional Parallelepipeds

The determinant is the essential algebraic tool which provides a way to give a unified treatment of the concept of p dimensional volume of a parallelepiped in \mathbb{R}^M . Here is the definition of what is meant by such a thing.

Definition 20.3.1 Let $\mathbf{u}_1, \dots, \mathbf{u}_p$ be vectors in $\mathbb{R}^M, M \geq p$. The parallelepiped determined by these vectors will be denoted by $P(\mathbf{u}_1, \dots, \mathbf{u}_p)$ and it is defined as

$$P(\mathbf{u}_1, \dots, \mathbf{u}_p) \equiv \left\{ \sum_{j=1}^p s_j \mathbf{u}_j : s_j \in [0, 1] \right\} = UQ, \quad Q = [0, 1]^p$$

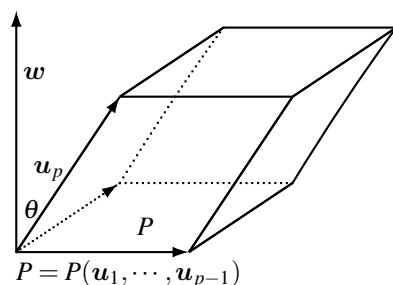
where $U = (\mathbf{u}_1 \quad \cdots \quad \mathbf{u}_p)$. The volume of this parallelepiped is defined as

$$\text{volume of } P(\mathbf{u}_1, \dots, \mathbf{u}_p) \equiv v(P(\mathbf{u}_1, \dots, \mathbf{u}_p)) \equiv (\det(G))^{1/2}.$$

where $G_{ij} = \mathbf{u}_i \cdot \mathbf{u}_j$. This $G = U^T U$ is called the metric tensor. If the vectors \mathbf{u}_i are dependent, this definition will give the volume to be 0.

First let's observe the last assertion is true. Say $\mathbf{u}_i = \sum_{j \neq i} \alpha_j \mathbf{u}_j$. Then the i^{th} row of G is a linear combination of the other rows using the scalars α_j and so from the properties of the determinant, the determinant of this matrix is indeed zero as it should be. Indeed, $\mathbf{u}_i \cdot \mathbf{u}_k = \sum_{j \neq i} \alpha_j \mathbf{u}_j \cdot \mathbf{u}_k$.

A parallelepiped is a sort of a squashed box. Here is a picture which shows



the relationship between $P(u_1, \dots, u_{p-1})$ and $P(u_1, \dots, u_p)$. In a sense, we can define the volume any way desired, but if it is to be reasonable, the following relationship must hold. The appropriate definition of the volume of $P(u_1, \dots, u_p)$ in terms of $P(u_1, \dots, u_{p-1})$ is $v(P(u_1, \dots, u_p)) =$

$$|u_p \cdot w| v(P(u_1, \dots, u_{p-1})) \quad (20.13)$$

where w is any unit vector perpendicular to each of u_1, \dots, u_{p-1} . Note $|u_p \cdot w| = |u_p| |\cos \theta|$ from the geometric meaning of the dot product. In the case where $p = 1$, the parallelepiped $P(v)$ consists of the single vector and the one dimensional volume should be $|v| = (v^T v)^{1/2} = (v \cdot v)^{1/2}$. Now having made this definition, I will show that $\det(G)^{1/2}$ is the appropriate definition of $v(P(u_1, \dots, u_p))$ for every p .

As just pointed out, this is the only reasonable definition of volume in the case of one vector. The next theorem shows that it is the only reasonable definition of volume of a parallelepiped in the case of p vectors because 20.13 holds.

Theorem 20.3.2 *If we desire 20.13 to hold for any w perpendicular to each u_i , then we obtain the definition of 20.3.1 for $v(P(u_1, \dots, u_p))$ in terms of determinants.*

Proof: So assume we want 20.13 to hold. Suppose the determinant formula holds for $P(u_1, \dots, u_{p-1})$. It is necessary to show that if w is a unit vector perpendicular to each u_1, \dots, u_{p-1} then $|u_p \cdot w| v(P(u_1, \dots, u_{p-1}))$ reduces to $\det(G)^{1/2}$. By the Gram Schmidt procedure there is (w_1, \dots, w_p) an orthonormal basis for $\text{span}(u_1, \dots, u_p)$ such that $\text{span}(w_1, \dots, w_k) = \text{span}(u_1, \dots, u_k)$ for each $k \leq p$. We can pick $w_p = w$ the given unit vector perpendicular to each u_i . First note that since $\{w_k\}_{k=1}^p$ is an orthonormal basis for $\text{span}(u_1, \dots, u_p)$,

$$u_j = \sum_{k=1}^p (u_j \cdot w_k) w_k, \quad u_j \cdot u_i = \sum_{k=1}^p (u_j \cdot w_k) (u_i \cdot w_k)$$

Therefore, the ij^{th} entry of the $p \times p$ matrix $U^T U$ is just

$$(U^T U)_{ij} = \sum_{r=1}^p (u_i \cdot w_r) (w_r \cdot u_j)$$

which is the product of a $p \times p$ matrix M whose rj^{th} entry is $w_r \cdot u_j$ with its transpose. The vector w_p is a unit vector perpendicular to each u_j for $j \leq p-1$ so $w_p \cdot u_j = 0$ if $j < p$.

Now consider the vector

$$N \equiv \det \begin{pmatrix} w_1 & \cdots & w_{p-1} & \overset{=0}{w_p} \\ u_1 \cdot w_1 & \cdots & u_1 \cdot w_{p-1} & u_1 \cdot w_p \\ \vdots & & \vdots & \vdots \\ u_{p-1} \cdot w_1 & \cdots & u_{p-1} \cdot w_{p-1} & \overset{=0}{u_{p-1} \cdot w_p} \end{pmatrix}$$

which results from formally expanding along the top row. Note you would get the same thing expanding along the last column because as just noted, the last column on the right

is 0 except for the top entry, so every cofactor A_{1k} for the $1k^{th}$ position is \pm a determinant which has a column of zeros. Thus \mathbf{N} is a multiple of \mathbf{w}_p . Hence, for $j < p$, $\mathbf{N} \cdot \mathbf{u}_j = 0$. From what was just discussed and induction, $v(P(\mathbf{u}_1, \dots, \mathbf{u}_{p-1})) = \pm A_{1p} = \mathbf{N} \cdot \mathbf{w}_p$. Also $\mathbf{N} \cdot \mathbf{u}_p$ equals

$$\det \begin{pmatrix} \mathbf{u}_p \cdot \mathbf{w}_1 & \cdots & \mathbf{u}_p \cdot \mathbf{w}_{p-1} & \mathbf{u}_p \cdot \mathbf{w}_p \\ \mathbf{u}_1 \cdot \mathbf{w}_1 & \cdots & \mathbf{u}_1 \cdot \mathbf{w}_{p-1} & \mathbf{u}_1 \cdot \mathbf{w}_p \\ \vdots & & \vdots & \vdots \\ \mathbf{u}_{p-1} \cdot \mathbf{w}_1 & \cdots & \mathbf{u}_{p-1} \cdot \mathbf{w}_{p-1} & \mathbf{u}_{p-1} \cdot \mathbf{w}_p \end{pmatrix} = \pm \det(M)$$

Thus from induction and expanding along the last column,

$$\begin{aligned} |\mathbf{u}_p \cdot \mathbf{w}_p| v(P(\mathbf{u}_1, \dots, \mathbf{u}_{p-1})) &= |\mathbf{N} \cdot \mathbf{u}_p| = \det(M^T M)^{1/2} \\ &= \det(U^T U)^{1/2} = \det(G)^{1/2}. \end{aligned}$$

Now $\mathbf{w}_p = \mathbf{w}$ the unit vector perpendicular to each \mathbf{u}_j for $j \leq p-1$. Thus if 20.13, then the claimed determinant identity holds. ■

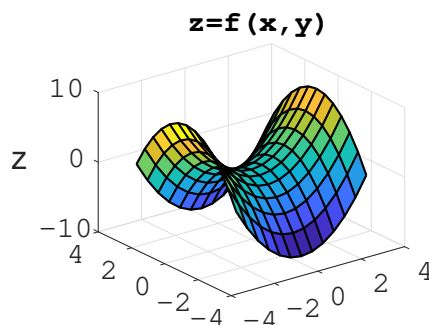
The theorem shows that the only reasonable definition of p dimensional volume of a parallelepiped is the one given in the above definition. Recall that these vectors are in \mathbb{R}^M . What is the role of \mathbb{R}^M ? It is just to provide an inner product. That is its only function. If $p = M$, then $\det(U^T U) = \det(U^T) \det(U) = \det(U)^2$ and so $\det(G)^{1/2} = |\det(U)|$.

Chapter 21

Functions of Many Variables

21.1 Graphs

In general, you really can't graph functions of many variables because we see in three dimensions. If you have a function of three variables, you would need four dimensions to graph it. However, in the case that $z = f(x, y)$ a scalar valued function of two variables, you can do so fairly well, especially with a computer algebra system. You graph $y \rightarrow f(x, y)$ for many values of x and $x \rightarrow f(x, y)$ for many values of y . This will result in a nice picture of a surface. For example, consider the graph of $z = f(x, y)$ where $f(x, y) = x^2 - y^2$.



To use MATLAB, to draw such a graph, modify the following syntax which was used for the above problem. Remember to get to a new line, you type shift enter.

```
[x,y]=meshgrid(-3:.5:3,-3:.5:3);  
z=x.^2-y.^2; surf(x,y,z,'LineWidth',2)
```

21.2 Review of Limits

Recall the concept of limit of a function of many variables. When $f : D(f) \rightarrow \mathbb{R}^q$ one can only consider in a meaningful way limits at limit points of the set $D(f)$.

Definition 21.2.1 Let A denote a nonempty subset of \mathbb{R}^p . A point \mathbf{x} is said to be a *limit point* of the set A if for every $r > 0$, $B(\mathbf{x}, r)$ contains infinitely many points of A .

Example 21.2.2 Let S denote the set $\{(x, y, z) \in \mathbb{R}^3 : x, y, z \text{ are all in } \mathbb{N}\}$. Which points are limit points?

This set does not have any because any two of these points are at least as far apart as 1. Therefore, if \mathbf{x} is any point of \mathbb{R}^3 , $B(\mathbf{x}, 1/4)$ contains at most one point.

Example 21.2.3 Let U be an open set in \mathbb{R}^3 . Which points of U are limit points of U ?

They all are. From the definition of U being open, if $\mathbf{x} \in U$, There exists $B(\mathbf{x}, r) \subseteq U$ for some $r > 0$. Now consider the line segment $\mathbf{x} + t\mathbf{e}_1$ where $t \in [0, 1/2]$. This describes infinitely many points and they are all in $B(\mathbf{x}, r)$ because $|\mathbf{x} + t\mathbf{e}_1 - \mathbf{x}| = tr < r$. Therefore, every point of U is a limit point of U .

The case where U is open will be the one of most interest, but many other sets have limit points.

Definition 21.2.4 Let $\mathbf{f} : D(\mathbf{f}) \subseteq \mathbb{R}^p \rightarrow \mathbb{R}^q$ where $q, p \geq 1$ be a function and let \mathbf{x} be a limit point of $D(\mathbf{f})$. Then

$$\lim_{\mathbf{y} \rightarrow \mathbf{x}} \mathbf{f}(\mathbf{y}) = \mathbf{L}$$

if and only if the following condition holds. For all $\varepsilon > 0$ there exists $\delta > 0$ such that if

$$0 < |\mathbf{y} - \mathbf{x}| < \delta \text{ and } \mathbf{y} \in D(\mathbf{f})$$

then,

$$|\mathbf{L} - \mathbf{f}(\mathbf{y})| < \varepsilon.$$

The condition that \mathbf{x} must be a limit point of $D(\mathbf{f})$ if you are to take a limit at \mathbf{x} is what makes the limit well defined.

Proposition 21.2.5 Let $\mathbf{f} : D(\mathbf{f}) \subseteq \mathbb{R}^p \rightarrow \mathbb{R}^q$ where $q, p \geq 1$ be a function and let \mathbf{x} be a limit point of $D(\mathbf{f})$. Then if $\lim_{\mathbf{y} \rightarrow \mathbf{x}} \mathbf{f}(\mathbf{y})$ exists, it must be unique.

Proof: Suppose $\lim_{\mathbf{y} \rightarrow \mathbf{x}} \mathbf{f}(\mathbf{y}) = \mathbf{L}_1$ and $\lim_{\mathbf{y} \rightarrow \mathbf{x}} \mathbf{f}(\mathbf{y}) = \mathbf{L}_2$. Then for $\varepsilon > 0$ given, let $\delta_i > 0$ correspond to \mathbf{L}_i in the definition of the limit and let $\delta = \min(\delta_1, \delta_2)$. Since \mathbf{x} is a limit point, there exists $\mathbf{y} \in B(\mathbf{x}, \delta) \cap D(\mathbf{f})$. Therefore,

$$|\mathbf{L}_1 - \mathbf{L}_2| \leq |\mathbf{L}_1 - \mathbf{f}(\mathbf{y})| + |\mathbf{f}(\mathbf{y}) - \mathbf{L}_2| < \varepsilon + \varepsilon = 2\varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, this shows $\mathbf{L}_1 = \mathbf{L}_2$. ■

The following theorem summarized many important interactions involving continuity. Most of this theorem has been proved in Theorem 15.8.6 on Page 330.

Theorem 21.2.6 Suppose \mathbf{x} is a limit point of $D(\mathbf{f})$ and

$$\lim_{\mathbf{y} \rightarrow \mathbf{x}} \mathbf{f}(\mathbf{y}) = \mathbf{L}, \quad \lim_{\mathbf{y} \rightarrow \mathbf{x}} \mathbf{g}(\mathbf{y}) = \mathbf{K}$$

where \mathbf{K} and \mathbf{L} are vectors in \mathbb{R}^p for $p \geq 1$. Then if $a, b \in \mathbb{R}$,

$$\lim_{\mathbf{y} \rightarrow \mathbf{x}} a\mathbf{f}(\mathbf{y}) + b\mathbf{g}(\mathbf{y}) = a\mathbf{L} + b\mathbf{K}, \tag{21.1}$$

$$\lim_{\mathbf{y} \rightarrow \mathbf{x}} \mathbf{f} \cdot \mathbf{g}(\mathbf{y}) = \mathbf{L} \cdot \mathbf{K} \tag{21.2}$$

Also, if h is a continuous function defined near L , then

$$\lim_{y \rightarrow x} h \circ f(y) = h(L). \quad (21.3)$$

For a vector valued function

$$f(y) = (f_1(y), \dots, f_q(y))^T,$$

$\lim_{y \rightarrow x} f(y) = L = (L_1, \dots, L_k)^T$ if and only if

$$\lim_{y \rightarrow x} f_k(y) = L_k \quad (21.4)$$

for each $k = 1, \dots, p$.

In the case where f and g have values in \mathbb{R}^3

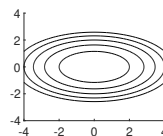
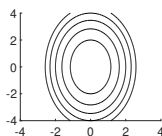
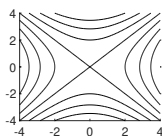
$$\lim_{y \rightarrow x} f(y) \times g(y) = L \times K. \quad (21.5)$$

Also recall Theorem 15.8.7 on Page 331.

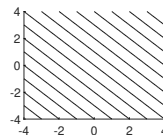
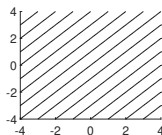
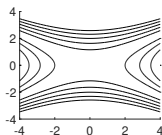
Theorem 21.2.7 For $f : D(f) \rightarrow \mathbb{R}^q$ and $x \in D(f)$ such that x is a limit point of $D(f)$, it follows f is continuous at x if and only if $\lim_{y \rightarrow x} f(y) = f(x)$.

21.3 Exercises

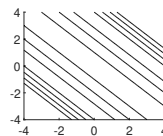
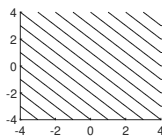
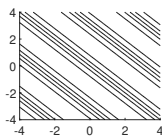
- Sketch the contour graph for $f(x, y) = (x-1)^2 + (y-2)^2$. This means you graph the relation $f(x, y) = c$ for various values of c . In this case, you would be graphing concentric circles with center at $(1, 2)$.
- Which of the following functions could correspond to the following contour graphs?
 $z = x^2 + 3y^2, z = 3x^2 + y^2, z = x^2 - y^2, z = x + y$.



- Which of the following functions could correspond to the following contour graphs?
 $z = x^2 - 3y^2, z = y^2 + 3x^2, z = x - y, z = x + y$.



- Which of the following functions could correspond to the following contour graphs?
 $z = \sin(x + y), z = x + y, z = (x + y)^2, z = x^2 - y$.



- Find the following limits if they exist. If they do not exist, explain why.

- (a) $\lim_{(x,y) \rightarrow (0,0)} \frac{x^2 - y^2}{x^2 + y^2}$ (d) $\lim_{(x,y) \rightarrow (0,0)} \frac{\sin(x^2 + 2y^2)}{x^2 + 2y^2}$
- (b) $\lim_{(x,y) \rightarrow (0,0)} \frac{2x^3 + xy^2 - x^2 - 2y^2}{x^2 + 2y^2}$ (e) $\lim_{(x,y) \rightarrow (0,0)} \frac{\sin(x^2 + 2y^2)}{2x^2 + y^2}$
- (c) $\lim_{(x,y) \rightarrow (0,0)} \frac{\sin(x^2 + y^2)}{x^2 + y^2}$ (f) $\lim_{(x,y) \rightarrow (0,0)} \frac{(x^2 - y^4)^2}{(x^2 + y^4)^2}$

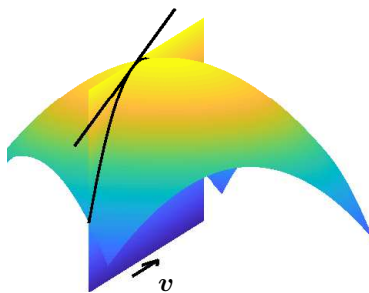
6. Find the following limits if they exist. If they do not exist, tell why.

- (a) $\lim_{(x,y) \rightarrow (0,0)} x \frac{(x^2 - y^4)^2}{(x^2 + y^4)^2}$
- (b) $\lim_{(x,y) \rightarrow (0,0)} \frac{x \sin(x^2 + 2y^2)}{2x^2 + y^2}$
- (c) $\lim_{(x,y) \rightarrow (0,0)} \frac{xy}{x^2 + y^2}$
- (d) $\lim_{(x,y) \rightarrow (1,0)} \frac{x^3 - 3x^2 + 3x - 1 - y^2x + y^2}{x^2 - 2x + 1 + y^2}$
7. *Suppose f is a function defined on a set D and that $a \in D$ is not a limit point of D . Show that if I define the notion of limit in the same way as above, then $\lim_{x \rightarrow a} f(x) = 5$. Show that it is also the case that $\lim_{x \rightarrow a} f(x) = 7$. In other words, the concept of limit is totally meaningless. This is why the insistence that the point a be a limit point of D .
8. *Show that the definition of continuity at $a \in D(f)$ is not dependent on a being a limit point of $D(f)$. The concept of limit and the concept of continuity are related at those points a which are limit points of the domain.

21.4 Directional and Partial Derivatives

21.4.1 The Directional Derivative

The directional derivative is just what its name suggests. It is the derivative of a function in a particular direction. The following picture illustrates the situation in the case of a function of two variables.



In this picture, $v \equiv (v_1, v_2)$ is a unit vector shown in the xy plane and $x_0 \equiv (x_0, y_0)$ is a point in the xy plane with $(x_0, y_0, f(x_0, y_0))$ being the point on the surface where there is a tangent line. When (x, y) moves in the direction of v , this results in a change in $z = f(x, y)$.

The directional derivative in this direction is the slope of the tangent line shown in the picture defined as

$$\lim_{t \rightarrow 0} \frac{f(x_0 + tv_1, y_0 + tv_2) - f(x_0, y_0)}{t}.$$

It tells how fast z is changing in this direction. A simple example of this is a person climbing a mountain. He could go various directions, some steeper than others. The directional derivative is just a measure of the steepness in a given direction. This motivates the following general definition of the directional derivative when it is not possible to draw pictures.

Definition 21.4.1 Let $f : U \rightarrow \mathbb{R}$ where U is an open set in \mathbb{R}^n and let \mathbf{v} be a unit vector. For $\mathbf{x} \in U$, define the **directional derivative** of f in the direction \mathbf{v} , at the point \mathbf{x} as

$$D_{\mathbf{v}}f(\mathbf{x}) \equiv \lim_{t \rightarrow 0} \frac{f(\mathbf{x} + t\mathbf{v}) - f(\mathbf{x})}{t}.$$

Example 21.4.2 Find the directional derivative of the function $f(x, y) = x^2y$ in the direction of $\mathbf{i} + \mathbf{j}$ at the point $(1, 2)$.

First you need a unit vector which has the same direction as the given vector. This unit vector is $\mathbf{v} \equiv \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$. Then to find the directional derivative from the definition, write the difference quotient described above. Thus $f(\mathbf{x} + t\mathbf{v}) = \left(1 + \frac{t}{\sqrt{2}}\right)^2 \left(2 + \frac{t}{\sqrt{2}}\right)$ and $f(\mathbf{x}) = 2$. Therefore,

$$\frac{f(\mathbf{x} + t\mathbf{v}) - f(\mathbf{x})}{t} = \frac{\left(1 + \frac{t}{\sqrt{2}}\right)^2 \left(2 + \frac{t}{\sqrt{2}}\right) - 2}{t},$$

and to find the directional derivative, you take the limit of this as $t \rightarrow 0$. However, this difference quotient equals $\frac{1}{4}\sqrt{2}(10 + 4t\sqrt{2} + t^2)$ and so, letting $t \rightarrow 0$, $D_{\mathbf{v}}f(1, 2) = \left(\frac{5}{2}\sqrt{2}\right)$.

There is something you must keep in mind about this. The direction vector must always be a unit vector¹.

21.4.2 Partial Derivatives

There are some special unit vectors which come to mind immediately. These are the vectors \mathbf{e}_i where $\mathbf{e}_i = (0, \dots, 0, 1, 0, \dots, 0)^T$ and the 1 is in the i^{th} position. The partial derivatives are simply directional derivatives taken in these special directions.

Definition 21.4.3 Let U be an open subset of \mathbb{R}^n and let $f : U \rightarrow \mathbb{R}$. Then letting $\mathbf{x} = (x_1, \dots, x_n)^T$ be a typical element of \mathbb{R}^n , $\frac{\partial f}{\partial x_i}(\mathbf{x}) \equiv D_{\mathbf{e}_i}f(\mathbf{x})$. This is called the **partial**

¹Actually, there is a more general formulation of the notion of directional derivative known as the Gateaux derivative in which the length of \mathbf{v} is not one but it is not considered here. This is actually a fairly old concept since Euler and Lagrange used something like it in their treatment of necessary conditions for the calculus of variations. The modern formulation is named after Gateaux who was killed in World War 1. This war killed some 40 million people. When sickness and disease and famine are included, the figure is some 80 million. One out of 20 French were killed. French soldiers died at the rate of about 900 per day. It is hard to find a reason for this conflict which would justify such an appalling loss of life.

derivative of f . Thus,

$$\begin{aligned}\frac{\partial f}{\partial x_i}(\mathbf{x}) &\equiv \lim_{t \rightarrow 0} \frac{f(\mathbf{x} + t\mathbf{e}_i) - f(\mathbf{x})}{t} \\ &= \lim_{t \rightarrow 0} \frac{f(x_1, \dots, x_i + t, \dots, x_n) - f(x_1, \dots, x_i, \dots, x_n)}{t},\end{aligned}$$

and to find the partial derivative, differentiate with respect to the variable of interest and regard all the others as constants. Other notation for this partial derivative is f_{x_i} , $f_{,i}$, or $D_i f$. If $y = f(\mathbf{x})$, the partial derivative of f with respect to x_i may also be denoted by $\frac{\partial y}{\partial x_i}$ or y_{x_i} or $D_{x_i} f$.

Example 21.4.4 Find $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$, and $\frac{\partial f}{\partial z}$ if $f(x, y) = y \sin x + x^2 y + z$.

From the definition above, $\frac{\partial f}{\partial x} = y \cos x + 2xy$, $\frac{\partial f}{\partial y} = \sin x + x^2$, and $\frac{\partial f}{\partial z} = 1$. Having taken one partial derivative, there is no reason to stop doing it. Thus, one could take the partial derivative with respect to y of the partial derivative with respect to x , denoted by $\frac{\partial^2 f}{\partial y \partial x}$ or f_{xy} . In the above example, $\frac{\partial^2 f}{\partial y \partial x} = f_{xy} = \cos x + 2x$. Also observe that $\frac{\partial^2 f}{\partial x \partial y} = f_{yx} = \cos x + 2x$.

Higher order partial derivatives are defined by analogy to the above. Thus in the above example,

$$f_{yxx} = -\sin x + 2.$$

These partial derivatives, f_{xy} are called mixed partial derivatives.

There is an interesting relationship between the directional derivatives and the partial derivatives under suitable conditions described later.

Definition 21.4.5 Suppose $f : U \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$ where U is an open set and the partial derivatives of f all exist. Define the **gradient** of f denoted $\nabla f(\mathbf{x})$ to be the vector

$$\nabla f(\mathbf{x}) = (f_{x_1}(\mathbf{x}), f_{x_2}(\mathbf{x}), \dots, f_{x_n}(\mathbf{x}))^T.$$

Proposition 21.4.6 In the situation of Definition 21.4.5, if the partial derivatives are continuous, then for \mathbf{v} a unit vector $D_{\mathbf{v}} f(\mathbf{x}) = \nabla f(\mathbf{x}) \cdot \mathbf{v}$.

This proposition will be proved in a more general setting later. For now, you can use it to compute directional derivatives.

Example 21.4.7 Find the directional derivative of the function

$$f(x, y) = \sin(2x^2 + y^3)$$

at $(1, 1)$ in the direction $\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)^T$.

First find the gradient. $\nabla f(x, y) = (4x \cos(2x^2 + y^3), 3y^2 \cos(2x^2 + y^3))^T$. Therefore, $\nabla f(1, 1) = (4 \cos(3), 3 \cos(3))^T$. The directional derivative is therefore,

$$(4 \cos(3), 3 \cos(3))^T \cdot \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)^T = \frac{7}{2} (\cos 3) \sqrt{2}.$$

Another important observation is that the gradient gives the direction in which the function changes most rapidly. The following proposition will be proved later.

Proposition 21.4.8 *In the situation of Definition 21.4.5, suppose $\nabla f(\mathbf{x}) \neq \mathbf{0}$. Then the direction in which f increases most rapidly, that is the direction in which the directional derivative is largest, is the direction of the gradient. Thus $\mathbf{v} = \nabla f(\mathbf{x}) / |\nabla f(\mathbf{x})|$ is the unit vector which maximizes $D_{\mathbf{v}}f(\mathbf{x})$ and this maximum value is $|\nabla f(\mathbf{x})|$. Similarly, $\mathbf{v} = -\nabla f(\mathbf{x}) / |\nabla f(\mathbf{x})|$ is the unit vector which minimizes $D_{\mathbf{v}}f(\mathbf{x})$ and this minimum value is $-|\nabla f(\mathbf{x})|$.*

The concept of a **directional derivative for a vector valued function** is also easy to define although the geometric significance expressed in pictures is not.

Definition 21.4.9 *Let $\mathbf{f} : U \rightarrow \mathbb{R}^p$ where U is an open set in \mathbb{R}^n and let \mathbf{v} be a unit vector. For $\mathbf{x} \in U$, define the directional derivative of \mathbf{f} in the direction \mathbf{v} , at the point \mathbf{x} as*

$$D_{\mathbf{v}}\mathbf{f}(\mathbf{x}) \equiv \lim_{t \rightarrow 0} \frac{\mathbf{f}(\mathbf{x} + t\mathbf{v}) - \mathbf{f}(\mathbf{x})}{t}.$$

Example 21.4.10 *Let $\mathbf{f}(x, y) = (xy^2, yx)^T$. Find the directional derivative in the direction $(1, 2)^T$ at the point (x, y) .*

First, a unit vector in this direction is $(1/\sqrt{5}, 2/\sqrt{5})^T$ and from the definition, the desired limit is

$$\begin{aligned} \lim_{t \rightarrow 0} \frac{\left((x+t(1/\sqrt{5})) (y+t(2/\sqrt{5}))^2 - xy^2, (x+t(1/\sqrt{5})) (y+t(2/\sqrt{5})) - xy \right)}{t} \\ = \lim_{t \rightarrow 0} \left(\frac{4}{5}xy\sqrt{5} + \frac{4}{5}xt + \frac{1}{5}\sqrt{5}y^2 + \frac{4}{5}ty + \frac{4}{25}t^2\sqrt{5}, \frac{2}{5}x\sqrt{5} + \frac{1}{5}y\sqrt{5} + \frac{2}{5}t \right) \\ = \left(\frac{4}{5}xy\sqrt{5} + \frac{1}{5}\sqrt{5}y^2, \frac{2}{5}x\sqrt{5} + \frac{1}{5}y\sqrt{5} \right). \end{aligned}$$

You see from this example and the above definition that all you have to do is to form the vector which is obtained by replacing each component of the vector with its directional derivative. In particular, you can take partial derivatives of vector valued functions and use the same notation.

Example 21.4.11 *Find the partial derivative with respect to x of the function $\mathbf{f}(x, y, z, w) = (xy^2, z \sin(xy), z^3x)^T$.*

From the above definition, $\mathbf{f}_x(x, y, z) = D_1\mathbf{f}(x, y, z) = (y^2, zy \cos(xy), z^3)^T$.

21.5 Exercises

1. Find the directional derivative of $f(x, y, z) = x^2y + z^4$ in the direction of the vector $(1, 3, -1)$ when $(x, y, z) = (1, 1, 1)$.
2. Find the directional derivative of $f(x, y, z) = \sin(x + y^2) + z$ in the direction of the vector $(1, 2, -1)$ when $(x, y, z) = (1, 1, 1)$.

3. Find the directional derivative of $f(x, y, z) = \ln(x + y^2) + z^2$ in the direction of the vector $(1, 1, -1)$ when $(x, y, z) = (1, 1, 1)$.
4. Using the conclusion of Proposition 21.4.6, prove Proposition 21.4.8 from the geometric description of the dot product, the one which says the dot product is the product of the lengths of the vectors and the cosine of the included angle which is no larger than π .
5. Find the largest value of the directional derivative of $f(x, y, z) = \ln(x + y^2) + z^2$ at the point $(1, 1, 1)$.
6. Find the smallest value of the directional derivative of $f(x, y, z) = x \sin(4xy^2) + z^2$ at the point $(1, 1, 1)$.
7. An ant falls to the top of a stove having temperature $T(x, y) = x^2 \sin(x + y)$ at the point $(2, 3)$. In what direction should the ant go to minimize the temperature? In what direction should he go to maximize the temperature?
8. Find the partial derivative with respect to y of the function $f(x, y, z, w) = (y^2, z^2 \sin(xy), z^3 x)^T$.
9. Find the partial derivative with respect to x of the function $f(x, y, z, w) = (wx, zx \sin(xy), z^3 x)^T$.
10. Find $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$, and $\frac{\partial f}{\partial z}$ for $f =$
 - (a) $x^2 y^2 z + w$
 - (b) $e^2 + xy + z^2$
 - (c) $\sin(z^2) + \cos(xy)$
 - (d) $\ln(x^2 + y^2 + 1) + e^z$
 - (e) $\sin(xyz) + \cos(xy)$
11. Find $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$, and $\frac{\partial f}{\partial z}$ for $f =$
 - (a) $x^2 y + \cos(xy) + z^3 y$
 - (b) $e^{x^2 + y^2} z \sin(x + y)$
 - (c) $z^2 \sin^3(e^{x^2 + y^3})$
 - (d) $x^2 \cos(\sin(\tan(z^2 + y^2)))$
 - (e) $x^{y^2 + z}$
12. Suppose

$$f(x, y) = \begin{cases} \frac{2xy + 6x^3 + 12xy^2 + 18yx^2 + 36y^3 + \sin(x^3) + \tan(3y^3)}{3x^2 + 6y^2} & \text{if } (x, y) \neq (0, 0) \\ 0 & \text{if } (x, y) = (0, 0). \end{cases}$$
 Find $\frac{\partial f}{\partial x}(0, 0)$ and $\frac{\partial f}{\partial y}(0, 0)$.
13. Why must the vector in the definition of the directional derivative be a unit vector? **Hint:** Suppose not. Would the directional derivative be a correct manifestation of steepness?

21.6 Mixed Partial Derivatives

Under certain conditions the **mixed partial derivatives** will always be equal. This astonishing fact may have been known to Euler in 1734.²

Theorem 21.6.1 Suppose $f : U \subseteq \mathbb{R}^2 \rightarrow \mathbb{R}$ 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 $|t|, |s| < r/2$ and consider

$$\Delta(s, t) \equiv \frac{1}{st} \left\{ \overbrace{f(x+t, y+s) - f(x+t, y)}^{h(t)} - \overbrace{(f(x, y+s) - f(x, y))}^{h(0)} \right\}. \quad (21.6)$$

Note that $(x+t, y+s) \in U$ because

$$|(x+t, y+s) - (x, y)| = |(t, s)| = (t^2 + s^2)^{1/2} \leq \left(\frac{r^2}{4} + \frac{r^2}{4} \right)^{1/2} = \frac{r}{\sqrt{2}} < r.$$

As implied above, $h(t) \equiv f(x+t, y+s) - f(x, y+s)$. Therefore, by the mean value theorem from calculus and the (one variable) chain rule,

$$\begin{aligned} \Delta(s, t) &= \frac{1}{st} (h(t) - h(0)) = \frac{1}{st} h'(\alpha t) t \\ &= \frac{1}{s} (f_x(x + \alpha t, y+s) - f_x(x, y+s)) \end{aligned}$$

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 21.6, $\Delta(s, t)$ is also 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 $(s, t) \rightarrow (0, 0)$ and using the continuity of f_{xy} and f_{yx} at (x, y) ,

$$\lim_{(s, t) \rightarrow (0, 0)} \Delta(s, t) = f_{xy}(x, y) = f_{yx}(x, y). \blacksquare$$

The following is obtained from the above by simply fixing all the variables except for the two of interest.

²Leonhard Euler 15 April 1707 - 18 September 1783 was the most prolific mathematician ever to have lived. His contributions also included fundamental work in fluid mechanics and engineering. For example, the formula for the stiffness of a beam which involves a moment of inertia is due to him. He wrote about 30,000 pages. He even wrote on music and theology. With Lagrange, he invented calculus of variations in which one looks for an unknown function maximizing a functional.

Euler had the ability to do huge computations in his head. He also had a memory which allowed him to memorize entire works of literature such as the Aeneid. He is also remembered for his work in logic, number theory, and graph theory. The notation π and e are due to him as is Euler's formula discussed earlier.

He was a kind and generous man and a devout Christian who believed the Bible was inspired. For the last part of his life, he was essentially blind. They didn't know how to treat things like cataracts back then.

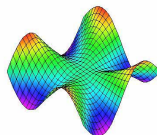
Corollary 21.6.2 Suppose U is an open subset of \mathbb{R}^n and $f : U \rightarrow \mathbb{R}$ has the property that for two indices k, l , f_{x_k} , f_{x_l} , $f_{x_l x_k}$, and $f_{x_k x_l}$ exist on U and $f_{x_k x_l}$ and $f_{x_l x_k}$ are both continuous at $\mathbf{x} \in U$. Then $f_{x_k x_l}(\mathbf{x}) = f_{x_l x_k}(\mathbf{x})$.

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 [3].

Example 21.6.3 Let

$$f(x, y) = \begin{cases} \frac{xy(x^2 - y^2)}{x^2 + y^2} & \text{if } (x, y) \neq (0, 0) \\ 0 & \text{if } (x, y) = (0, 0) \end{cases}$$

Here is a picture of the graph of this function. It looks innocuous but isn't.



From the definition of partial derivatives it follows immediately that $f_x(0, 0) = f_y(0, 0) = 0$. Using the standard rules of differentiation, for $(x, y) \neq (0, 0)$,

$$f_x = y \frac{x^4 - y^4 + 4x^2 y^2}{(x^2 + y^2)^2}, \quad f_y = x \frac{x^4 - y^4 - 4x^2 y^2}{(x^2 + y^2)^2}$$

Now

$$f_{xy}(0, 0) \equiv \lim_{y \rightarrow 0} \frac{f_x(0, y) - f_x(0, 0)}{y} = \lim_{y \rightarrow 0} \frac{-y^4}{(y^2)^2} = -1$$

while

$$f_{yx}(0, 0) \equiv \lim_{x \rightarrow 0} \frac{f_y(x, 0) - f_y(0, 0)}{x} = \lim_{x \rightarrow 0} \frac{x^4}{(x^2)^2} = 1$$

showing that, although the mixed partial derivatives do exist at $(0, 0)$, they are not equal there.

21.7 Partial Differential Equations

Partial differential equations are equations which involve the partial derivatives of some function. The most famous partial differential equations involve the **Laplacian**, named after Laplace³.

Definition 21.7.1 Let u be a function of n variables. Then

$$\Delta u \equiv \sum_{k=1}^n u_{x_k x_k}$$

This is also written as $\nabla^2 u$. The symbol Δ or ∇^2 is called the *Laplacian*. When $\Delta u = 0$ the function u is called **harmonic**. **Laplace's equation** is $\Delta u = 0$. The **heat equation** is $u_t - \Delta u = 0$ and the **wave equation** is $u_{tt} - \Delta u = 0$.

³Laplace was a great physicist and mathematician of the 1700's. He made fundamental contributions to mechanics and astronomy.

Example 21.7.2 Find the Laplacian of $u(x, y) = x^2 - y^2$.

$u_{xx} = 2$ while $u_{yy} = -2$. Therefore, $\Delta u = u_{xx} + u_{yy} = 2 - 2 = 0$. Thus this function is harmonic, $\Delta u = 0$.

Example 21.7.3 Find $u_t - \Delta u$ where $u(t, x, y) = e^{-t} \cos x$.

In this case, $u_t = -e^{-t} \cos x$ while $u_{yy} = 0$ and $u_{xx} = -e^{-t} \cos x$ therefore, $u_t - \Delta u = 0$ and so u solves the heat equation $u_t - \Delta u = 0$.

Example 21.7.4 Let $u(t, x) = \sin t \cos x$. Find $u_{tt} - \Delta u$.

In this case, $u_{tt} = -\sin t \cos x$ while $\Delta u = -\sin t \cos x$. Therefore, u is a solution of the wave equation $u_{tt} - \Delta u = 0$.

21.8 Exercises

- Find $f_x, f_y, f_z, f_{xy}, f_{yx}, f_{xz}, f_{zx}, f_{zy}, f_{yz}$ for the following. Verify the mixed partial derivatives are equal.
 - $x^2 y^3 z^4 + \sin(xyz)$
 - $\sin(xyz) + x^2 yz$
 - $z \ln |x^2 + y^2 + 1|$
 - $e^{x^2 + y^2 + z^2}$
 - $\tan(xyz)$
- Suppose f is a continuous function and $f : U \rightarrow \mathbb{R}$ where U is an open set and suppose that $\mathbf{x} \in U$ has the property that for all \mathbf{y} near \mathbf{x} , $f(\mathbf{x}) \leq f(\mathbf{y})$. Prove that if f has all of its partial derivatives at \mathbf{x} , then $f_{x_i}(\mathbf{x}) = 0$ for each x_i . **Hint:** This is just a repeat of the usual one variable theorem seen in beginning calculus. You just do this one variable argument for each variable to get the conclusion.
- As an important application of Problem 2 consider the following. Experiments are done at n times, t_1, t_2, \dots, 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.
- Show that if $v(x, y) = u(\alpha x, \beta y)$, then $v_x = \alpha u_x$ and $v_y = \beta u_y$. State and prove a generalization to any number of variables.
- Let f be a function which has continuous derivatives. Show that $u(t, x) = f(x - ct)$ solves the wave equation $u_{tt} - c^2 u_{xx} = 0$. What about $u(x, t) = f(x + ct)$?

6. D'Alembert found a formula for the solution to the wave equation $u_{tt} = c^2 u_{xx}$ along with the initial conditions $u(x, 0) = f(x)$, $u_t(x, 0) = g(x)$. Here is how he did it. He looked for a solution of the form $u(x, t) = h(x + ct) + k(x - ct)$ and then found h and k in terms of the given functions f and g . He ended up with something like

$$u(x, t) = \frac{1}{2c} \int_{x-ct}^{x+ct} g(r) dr + \frac{1}{2} (f(x+ct) + f(x-ct)).$$

Fill in the details.

7. Determine which of the following functions satisfy Laplace's equation.

- | | |
|---------------------------------|-----------------------------------|
| (a) $x^3 - 3xy^2$ | (e) $3x^2 - y^3 + 4xy$ |
| (b) $3x^2y - y^3$ | (f) $3x^2y - y^3 + 4y$ |
| (c) $x^3 - 3xy^2 + 2x^2 - 2y^2$ | (g) $x^3 - 3x^2y^2 + 2x^2 - 2y^2$ |
| (d) $3x^2y - y^3 + 4xy$ | |

8. Show that $z = \sqrt{x^2 + y^2}$ is a solution to $x \frac{\partial z}{\partial x} + y \frac{\partial z}{\partial y} = z$.
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 $u_t - \Delta u = 0$.
10. Show that if a, b are scalars and u, v are functions which satisfy Laplace's equation then $au + bv$ also satisfies Laplace's equation. Verify a similar statement for the heat and wave equations.
11. Show that $u(x, t) = \frac{1}{\sqrt{t}} e^{-\frac{1}{4c^2 t} x^2}$ solves the heat equation $u_t = c^2 u_{xx}$.

Chapter 22

Derivative of a Functions of Many Variables

Linear functions were just discussed. The derivative of a nonlinear function of many variables is a linear approximation to the function which is valid locally. You have $f : U \rightarrow \mathbb{R}^m$ where U is an open subset of \mathbb{R}^n and the derivative at some point is $T \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ such that near the point x , $f(x + v)$ is close to $T(v) + f(x)$. This is the main idea.

22.1 The Derivative of Functions of One Variable

First consider the notion of the derivative of a function of one variable.

Observation 22.1.1 Suppose a function f of one variable has a derivative at x . Then

$$\lim_{h \rightarrow 0} \frac{|f(x+h) - f(x) - f'(x)h|}{|h|} = 0.$$

This observation follows from the definition of the derivative of a function of one variable, namely

$$f'(x) \equiv \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}.$$

Thus

$$\lim_{h \rightarrow 0} \frac{|f(x+h) - f(x) - f'(x)h|}{|h|} = \lim_{h \rightarrow 0} \left| \frac{f(x+h) - f(x)}{h} - f'(x) \right| = 0$$

Definition 22.1.2 A vector valued function of a vector v is called $o(v)$ (referred to as “little o of v ”) if

$$\lim_{|v| \rightarrow 0} \frac{o(v)}{|v|} = 0. \quad (22.1)$$

Thus for a function of one variable, the function $f(x+h) - f(x) - f'(x)h$ is $o(h)$. When we say a function is $o(h)$, it is used like an adjective. It is like saying the function is white or black or green or fat or thin. The term is used very imprecisely. Thus in general,

$$o(v) = o(v) + o(v), o(v) = 45 \times o(v), o(v) = o(v) - o(v), \text{etc.}$$

When you add two functions with the property of the above definition, you get another one having that same property. When you multiply by 45, the property is also retained, as it is when you subtract two such functions. How could something so sloppy be useful? The notation is useful precisely because it prevents you from obsessing over things which are not relevant and should be ignored.

Theorem 22.1.3 *Let $f : (a, b) \rightarrow \mathbb{R}$ be a function of one variable. Then $f'(x)$ exists if and only if there exists p such that*

$$f(x+h) - f(x) = ph + o(h) \quad (22.2)$$

In this case, $p = f'(x)$.

Proof: From the above observation it follows that if $f'(x)$ does exist, then 22.2 holds. Suppose then that 22.2 is true. Then

$$\frac{f(x+h) - f(x)}{h} - p = \frac{o(h)}{h}.$$

Taking a limit, you see that

$$p = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}$$

and that in fact this limit exists which shows that $p = f'(x)$. ■

This theorem shows that one way to define $f'(x)$ is as the number p , if there is one, which has the property that

$$f(x+h) = f(x) + ph + o(h).$$

You should think of p as the linear transformation resulting from multiplication by the 1×1 matrix (p) .

Example 22.1.4 *Let $f(x) = x^3$. Find $f'(x)$.*

$$f(x+h) = (x+h)^3 = x^3 + 3x^2h + 3xh^2 + h^3 = f(x) + 3x^2h + (3xh + h^2)h.$$

Since $(3xh + h^2)h = o(h)$, it follows $f'(x) = 3x^2$.

Example 22.1.5 *Let $f(x) = \sin(x)$. Find $f'(x)$. $f(x+h) - f(x) =$*

$$\begin{aligned} \sin(x+h) - \sin(x) &= \sin(x)\cos(h) + \cos(x)\sin(h) - \sin(x) \\ &= \cos(x)\sin(h) + \sin(x)\frac{(\cos(h)-1)}{h}h \\ &= \cos(x)h + \cos(x)\frac{(\sin(h)-h)}{h}h + \sin(x)\frac{(\cos(h)-1)}{h}h. \end{aligned}$$

Now

$$\cos(x)\frac{(\sin(h)-h)}{h}h + \sin(x)\frac{(\cos(h)-1)}{h}h = o(h). \quad (22.3)$$

Remember the fundamental limits which allowed you to find the derivative of $\sin(x)$ were

$$\lim_{h \rightarrow 0} \frac{\sin(h)}{h} = 1, \quad \lim_{h \rightarrow 0} \frac{\cos(h) - 1}{h} = 0. \quad (22.4)$$

These same limits are what is needed to verify 22.3.

How can you tell whether a function of two variables (u, v) is $o\left(\begin{smallmatrix} u \\ v \end{smallmatrix}\right)$? In general, there is no substitute for the definition, but you can often identify this property by observing that the expression involves only “higher order terms”. These are terms like u^2v, uv, v^4 , etc. If you sum the exponents on the u and the v you get something larger than 1. For example,

$$\left| \frac{vu}{\sqrt{u^2 + v^2}} \right| \leq \frac{1}{2} (u^2 + v^2) \frac{1}{\sqrt{u^2 + v^2}} = \frac{1}{2} \sqrt{u^2 + v^2}$$

and this converges to 0 as $(u, v) \rightarrow (0, 0)$. This follows from the inequality $|uv| \leq \frac{1}{2} (u^2 + v^2)$ which you can verify from $(u - v)^2 \geq 0$. Similar considerations apply in higher dimensions also. In general, this is a hard question because it involves a limit of a function of many variables. Furthermore, there is really no substitute for answering this question, because its resolution involves the definition of whether a function is differentiable. That may be why we spend most of our time on one dimensional considerations which involve taking the partial derivatives. The following exercises should help give you an idea of how to determine whether something is o .

22.2 The Derivative

The way of thinking about the derivative in Theorem 22.1.3 is exactly what is needed to define the derivative of a function of n variables. One can argue that it is also the right way to define the derivative of a function of one variable in order to reduce confusion later on.

As observed by Deudonne, “...In the classical teaching of Calculus, this idea (that the derivative is a linear transformation) is immediately obscured by the accidental fact that, on a one-dimensional vector space, there is a one-to-one correspondence between linear forms and numbers, and therefore the derivative at a point is defined as a number instead of a linear form. This slavish subservience to the shibboleth¹ of numerical interpretation at any cost becomes much worse when dealing with functions of several variables...”

In fact, the derivative is a linear transformation and it is useless to pretend otherwise. This is the main reason for including the introductory material on linear algebra in this book.

Recall the following definition.

Definition 22.2.1 A function T which maps \mathbb{R}^n to \mathbb{R}^p is called a linear transformation if for every pair of scalars, a, b and vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, it follows that $T(a\mathbf{x} + b\mathbf{y}) = aT(\mathbf{x}) + bT(\mathbf{y})$.

¹In the Bible, there was a battle between Ephraimites and Gileadites during the time of Jephthah, the judge who sacrificed his daughter to Jehovah, one of several instances of human sacrifice in the Bible. The cause of this battle was very strange. However, the Ephraimites lost and when they tried to cross a river to get back home, they had to say shibboleth. If they said “shibboleth” they were killed because their inability to pronounce the “sh” sound identified them as Ephraimites. They usually don’t tell this story in Sunday school. The word has come to signify something which is arbitrary and no longer important.

Recall that from the properties of matrix multiplication, if A is a $p \times n$ matrix, and if \mathbf{x}, \mathbf{y} are vectors in \mathbb{R}^n , then $A(a\mathbf{x} + b\mathbf{y}) = aA(\mathbf{x}) + bA(\mathbf{y})$. Thus you can define a linear transformation by multiplying by a matrix. Of course the simplest example is that of a 1×1 matrix or number. You can think of the number 3 as a linear transformation T mapping \mathbb{R} to \mathbb{R} according to the rule $Tx = 3x$. It satisfies the properties needed for a linear transformation because $3(ax + by) = a3x + b3y = aTx + bTy$. The case of the derivative of a scalar valued function of one variable is of this sort. You get a number for the derivative. However, you can think of this number as a linear transformation and this is the way you must think of it for a function of n variables. First there is a useful lemma.

Lemma 22.2.2 *Let $T \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$. Then there is a constant C such that $|T\mathbf{x}| \leq C|\mathbf{v}|$.*

Proof: Let A be the matrix of T . Then, using the Cauchy Schwarz inequality Lemma 13.5.4, the following computation shows the desired result.

$$\begin{aligned} |T\mathbf{x}| &= |A\mathbf{x}| = \left(\sum_{j=1}^m |(A\mathbf{x})_j|^2 \right)^{1/2} = \left(\sum_{j=1}^m \left| \sum_{k=1}^n A_{jk}x_k \right|^2 \right)^{1/2} \\ &\leq \left(\sum_{j=1}^m \left| \left(\sum_{k=1}^n (A_{jk})^2 \right)^{1/2} \left(\sum_{k=1}^n |x_k|^2 \right)^{1/2} \right|^2 \right)^{1/2} \\ &= |\mathbf{x}| \left(\sum_{j=1}^m \left| \sum_{k=1}^n (A_{jk})^2 \right| \right)^{1/2} \blacksquare \end{aligned}$$

Definition 22.2.3 *Let $\mathbf{f} : U \rightarrow \mathbb{R}^p$ where U is an open set in \mathbb{R}^n for $n, p \geq 1$ and let $\mathbf{x} \in U$ be given. Then \mathbf{f} is defined to be **differentiable** at $\mathbf{x} \in U$ if and only if there exists a linear transformation T such that,*

$$\mathbf{f}(\mathbf{x} + \mathbf{h}) = \mathbf{f}(\mathbf{x}) + T\mathbf{h} + \mathbf{o}(\mathbf{h}). \quad (22.5)$$

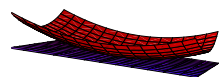
The derivative of the function \mathbf{f} , denoted by $D\mathbf{f}(\mathbf{x})$, is this linear transformation. Thus

$$\mathbf{f}(\mathbf{x} + \mathbf{h}) = \mathbf{f}(\mathbf{x}) + D\mathbf{f}(\mathbf{x})\mathbf{h} + \mathbf{o}(\mathbf{h})$$

If $\mathbf{h} = \mathbf{x} - \mathbf{x}_0$, this takes the form

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}(\mathbf{x}_0) + D\mathbf{f}(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + \mathbf{o}(\mathbf{x} - \mathbf{x}_0)$$

If you deleted the $\mathbf{o}(\mathbf{x} - \mathbf{x}_0)$ term and considered the function of \mathbf{x} given by what is left, this is called the linear approximation to the function at the point \mathbf{x}_0 . In the case where \mathbb{R} in \mathbb{R} one can draw a picture to illustrate this.



Of course the first and most obvious question is whether the linear transformation is unique. Otherwise, the definition of the derivative $D\mathbf{f}(\mathbf{x})$ would not be well defined.

Theorem 22.2.4 Suppose f is differentiable, as given above in 22.5. Then T is uniquely determined. Furthermore, the matrix of T is the following $p \times n$ matrix

$$\left(\begin{array}{ccc} \frac{\partial f(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f(\mathbf{x})}{\partial x_n} \end{array} \right)$$

where

$$\frac{\partial f}{\partial x_i}(\mathbf{x}) \equiv \lim_{h \rightarrow 0} \frac{f(\mathbf{x} + t\mathbf{e}_i) - f(\mathbf{x})}{t},$$

the k^{th} partial derivative of f .

Proof: Suppose T_1 is another linear transformation which works. Thus, letting t be a small positive real number,

$$f(\mathbf{x} + t\mathbf{h}) = f(\mathbf{x}) + Tt\mathbf{h} + o(t\mathbf{h}), \quad f(\mathbf{x} + t\mathbf{h}) = f(\mathbf{x}) + T_1t\mathbf{h} + o(t\mathbf{h})$$

Now $o(t\mathbf{h}) = o(t)$ and so, subtracting these yields $Tt\mathbf{h} - T_1t\mathbf{h} = o(t)$. Divide both sides by t to obtain $T\mathbf{h} - T_1\mathbf{h} = \frac{o(t)}{t}$. It follows on letting $t \rightarrow 0$ that $T\mathbf{h} = T_1\mathbf{h}$. Since \mathbf{h} is arbitrary, this shows that $T = T_1$. Thus the derivative is well defined. So what is the matrix of this linear transformation? From Theorem 18.1.6, this is the matrix whose i^{th} column is Te_i . However, from the definition of T , letting $t \neq 0$,

$$\frac{f(\mathbf{x} + t\mathbf{e}_i) - f(\mathbf{x})}{t} = \frac{1}{t} (T(te_i) + o(te_i)) = T(\mathbf{e}_i) + \frac{o(te_i)}{t} = T(\mathbf{e}_i) + \frac{o(t)}{t}$$

Then letting $t \rightarrow 0$, it follows that $Te_i = \frac{\partial f}{\partial x_i}(\mathbf{x})$. Recall from theorem 18.1.6 this shows that the matrix of the linear transformation is as claimed. ■

Other notations which are often used for this matrix or the linear transformation are $f'(\mathbf{x})$, $J(\mathbf{x})$, and even $\frac{\partial f}{\partial \mathbf{x}}$ or $\frac{df}{d\mathbf{x}}$. Also, the above definition can now be written in the form

$$f(\mathbf{x} + \mathbf{v}) = f(\mathbf{x}) + \sum_{j=1}^p \frac{\partial f(\mathbf{x})}{\partial x_j} v_j + o(\mathbf{v})$$

or

$$f(\mathbf{x} + \mathbf{v}) - f(\mathbf{x}) = \left(\begin{array}{ccc} \frac{\partial f(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f(\mathbf{x})}{\partial x_n} \end{array} \right) \mathbf{v} + o(\mathbf{v})$$

Here is an example of a scalar valued nonlinear function.

Example 22.2.5 Suppose $f(x, y) = \sqrt{xy}$. Find the approximate change in f if x goes from 1 to 1.01 and y goes from 4 to 3.99.

We can do this by noting that

$$\begin{aligned} f(1.01, 3.99) - f(1, 4) &\approx f_x(1, 2)(.01) + f_y(1, 2)(-.01) \\ &= 1(.01) + \frac{1}{4}(-.01) = 7.5 \times 10^{-3}. \end{aligned}$$

Of course the exact value is

$$\sqrt{(1.01)(3.99)} - \sqrt{4} = 7.4610831 \times 10^{-3}.$$

Notation 22.2.6 When f is a scalar valued function of n variables, the following is often written to express the idea that a small change in f due to small changes in the variables can be expressed in the form

$$df(\mathbf{x}) = f_{x_1}(\mathbf{x})dx_1 + \cdots + f_{x_n}(\mathbf{x})dx_n$$

where the small change in x_i is denoted as dx_i . As explained above, df is the approximate change in the function f . Sometimes df is referred to as the differential of f .

Let $\mathbf{f} : U \rightarrow \mathbb{R}^q$ where U is an open subset of \mathbb{R}^p and \mathbf{f} is differentiable. It was just shown that

$$\mathbf{f}(\mathbf{x} + \mathbf{v}) = \mathbf{f}(\mathbf{x}) + \left(\frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_1} \quad \cdots \quad \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_p} \right) \mathbf{v} + \mathbf{o}(\mathbf{v}).$$

Taking the i^{th} coordinate of the above equation yields

$$f_i(\mathbf{x} + \mathbf{v}) = f_i(\mathbf{x}) + \sum_{j=1}^p \frac{\partial f_i(\mathbf{x})}{\partial x_j} v_j + o(v),$$

and it follows that the term with a sum is nothing more than the i^{th} component of $J(\mathbf{x})\mathbf{v}$ where $J(\mathbf{x})$ is the $q \times p$ matrix

$$\begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_p} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_q}{\partial x_1} & \frac{\partial f_q}{\partial x_2} & \cdots & \frac{\partial f_q}{\partial x_p} \end{pmatrix}.$$

Thus

$$\mathbf{f}(\mathbf{x} + \mathbf{v}) = \mathbf{f}(\mathbf{x}) + J(\mathbf{x})\mathbf{v} + \mathbf{o}(\mathbf{v}), \quad (22.6)$$

and to reiterate, the linear transformation which results by multiplication by this $q \times p$ matrix is known as the derivative.

Sometimes x, y, z is written instead of x_1, x_2 , and x_3 . This is to save on notation and is easier to write and to look at although it lacks generality. When this is done it is understood that $x = x_1, y = x_2$, and $z = x_3$. Thus the derivative is the linear transformation determined by

$$\begin{pmatrix} f_{1x} & f_{1y} & f_{1z} \\ f_{2x} & f_{2y} & f_{2z} \\ f_{3x} & f_{3y} & f_{3z} \end{pmatrix}.$$

Example 22.2.7 Let A be a constant $m \times n$ matrix and consider $\mathbf{f}(\mathbf{x}) = A\mathbf{x}$. Find $D\mathbf{f}(\mathbf{x})$ if it exists.

$$\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) = A(\mathbf{x} + \mathbf{h}) - A\mathbf{x} = A\mathbf{h} = A\mathbf{h} + \mathbf{o}(\mathbf{h}).$$

In fact in this case, $\mathbf{o}(\mathbf{h}) = \mathbf{0}$. Therefore, $D\mathbf{f}(\mathbf{x}) = A$. Note that this looks the same as the case in one variable, $f(x) = ax$.

Example 22.2.8 Let $f(x, y, z) = xy + z^2x$. Find $Df(x, y, z)$.

Consider $f(x+h, y+k, z+l) - f(x, y, z)$. This is something which is easily computed from the definition of the function. It equals

$$(x+h)(y+k) + (z+l)^2(x+h) - (xy + z^2x)$$

Multiply everything together and collect the terms. This yields

$$(z^2 + y)h + xk + 2zxl + (hk + 2zlh + l^2x + l^2h)$$

It follows easily the last term at the end is $o(h, k, l)$ and so the derivative of this function is the linear transformation coming from multiplication by the matrix $((z^2 + y), x, 2zx)$ and so this is the derivative. It follows from this and the description of the derivative in terms of partial derivatives that

$$\frac{\partial f}{\partial x}(x, y, z) = z^2 + y, \quad \frac{\partial f}{\partial y}(x, y, z) = x, \quad \frac{\partial f}{\partial z}(x, y, z) = 2xz.$$

Of course you could compute these partial derivatives directly.

Given a function of many variables, how can you tell if it is differentiable? In other words, when you make the linear approximation, how can you tell easily that what is left over is $o(v)$. Sometimes you have to go directly to the definition and verify it is differentiable from the definition. For example, here is an interesting example of a function of one variable.

Example 22.2.9 Let $f(x) = \begin{cases} x^2 \sin(\frac{1}{x}) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$. Find $Df(0)$.

$$f(h) - f(0) = 0h + h^2 \sin\left(\frac{1}{h}\right) = o(h),$$

and so $Df(0) = 0$. If you find the derivative for $x \neq 0$, it is totally useless information if what you want is $Df(0)$. This is because the derivative turns out to be discontinuous. Try it. Find the derivative for $x \neq 0$ and try to obtain $Df(0)$ from it. You see, in this example you had to revert to the definition to find the derivative.

It isn't really too hard to use the definition even for more ordinary examples.

Example 22.2.10 Let $f(x, y) = \begin{pmatrix} x^2y + y^2 \\ y^3x \end{pmatrix}$. Find $Df(1, 2)$.

First of all, note that the thing you are after is a 2×2 matrix.

$$f(1, 2) = \begin{pmatrix} 6 \\ 8 \end{pmatrix}.$$

Then

$$f(1+h_1, 2+h_2) - f(1, 2) = \begin{pmatrix} (1+h_1)^2(2+h_2) + (2+h_2)^2 \\ (2+h_2)^3(1+h_1) \end{pmatrix} - \begin{pmatrix} 6 \\ 8 \end{pmatrix}$$

after some simplification,

$$= \begin{pmatrix} 4 & 5 \\ 8 & 12 \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} + \begin{pmatrix} 2h_1h_2 + 2h_1^2 + h_1^2h_2 + h_2^2 \\ 12h_1h_2 + 6h_2^2 + 6h_2^2h_1 + h_2^3 + h_2^3h_1 \end{pmatrix}$$

$$= \begin{pmatrix} 4 & 5 \\ 8 & 12 \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} + o(\mathbf{h}).$$

Therefore, the matrix of the derivative is $\begin{pmatrix} 4 & 5 \\ 8 & 12 \end{pmatrix}$. You let the $o(\mathbf{h})$ terms be the higher order terms, polynomials in the components of \mathbf{h} which have higher degree than 1.

Example 22.2.11 Let $\mathbf{f}(x, y) = \begin{pmatrix} x^3y + y^2 \\ xy^2 + 1 \end{pmatrix}$. Find $D\mathbf{f}(x, y)$.

Simple computations show that the matrix of this linear transformation is

$$D\mathbf{f}(x, y) = \begin{pmatrix} f_{1x}(x, y) & f_{1y}(x, y) \\ f_{2x}(x, y) & f_{2y}(x, y) \end{pmatrix} = \begin{pmatrix} 3x^2y & x^3 + 2y \\ y^2 & 2xy \end{pmatrix}$$

provided the function is differentiable. It is left as an exercise to verify that this does indeed serve as the derivative. A little later, a theorem is given which shows that, since the function is a C^1 function, it is indeed differentiable.

Example 22.2.12 Consider the open set O in the space of $p \times p$ matrices consisting of those which have an inverse. Let $\phi(A) \equiv \det(A)$. Then have a look at Problem 39 on Page 447 to see a description of $D\phi(F)$.

22.3 Exercises

1. Determine which of the following functions are $o(h)$.

- | | |
|--------------------------|-------------------------------------|
| (a) h^2 | (f) $\sin(h)$ |
| (b) $h \sin(h)$ | (g) $xh \sin(\sqrt{ h }) + x^5 h^2$ |
| (c) $ h ^{3/2} \ln(h)$ | (h) $\exp(-1/ h ^2)$ |
| (d) $h^2 x + y h^3$ | |
| (e) $\sin(h^2)$ | |

2. Here are some scalar valued functions of several variables. Determine which of these functions are $o(\mathbf{v})$. Here \mathbf{v} is a vector in \mathbb{R}^n , $\mathbf{v} = (v_1, \dots, v_n)$.

- | | |
|---------------------------|---|
| (a) $v_1 v_2$ | (e) $v_1(v_1 + v_2 + x v_3)$ |
| (b) $v_2 \sin(v_1)$ | (f) $(e^{v_1} - 1 - v_1)$ |
| (c) $v_1^2 + v_2$ | (g) $(\mathbf{x} \cdot \mathbf{v}) \mathbf{v} $ |
| (d) $v_2 \sin(v_1 + v_2)$ | |

3. Here are some vector valued functions of $\mathbf{v} \in \mathbb{R}^n$. Determine which ones are $o(\mathbf{v})$.

- | | |
|--|---|
| (a) $(\mathbf{x} \cdot \mathbf{v})\mathbf{v}$ | (d) $\sqrt{ (\mathbf{x} \cdot \mathbf{v}) } \mathbf{v} ^{1/2}$ |
| (b) $\sin(v_1)\mathbf{v}$ | (e) $\left(\sin\left(\sqrt{ \mathbf{x} \cdot \mathbf{v} }\right) - \sqrt{ \mathbf{x} \cdot \mathbf{v} }\right) \cdot \mathbf{v} ^{-1/4}$ |
| (c) $\sqrt{ (\mathbf{x} \cdot \mathbf{v}) } \mathbf{v} ^{2/3}$ | |

(f) $\exp(-1/|v|^2)$

(g) $v^T A v$ where A is an $n \times n$ matrix.

4. Show that if $f(x) = o(x)$, then $f'(0) = 0$.
5. Show that if $\lim_{h \rightarrow 0} f(x) = 0$ then $xf(x) = o(x)$.
6. Show that if $f'(0)$ exists and $f(0) = 0$, then $f(|x|^p) = o(x)$ whenever $p > 1$.
7. Use the definition of the derivative to find the 1×1 matrix which is the derivative of the following functions.
 - (a) $f(t) = t^2 + t$.
 - (b) $f(t) = t^3$.
 - (c) $f(t) = t \sin(t)$.
 - (d) $f(t) = \ln(t^2 + 1)$.
 - (e) $f(t) = t|t|$.
8. Show that if f is a real valued function defined on (a, b) and it achieves a local maximum at $x \in (a, b)$, then $Df(x) = 0$.
9. Use the above definition of the derivative to prove the product rule for functions of 1 variable.
10. Let $f(x, y) = x \sin(y)$. Compute the derivative directly from the definition.
11. Let $f(x, y) = x^2 \sin(y)$. Compute the derivative directly from the definition.
12. Let $f(x, y) = \begin{pmatrix} x^2 + y \\ y^2 \end{pmatrix}$. Compute the derivative directly from the definition.
13. Let $f(x, y) = \begin{pmatrix} x^2 y \\ x + y^2 \end{pmatrix}$. Compute the derivative directly from the definition.
14. Let $f(x, y) = x^\alpha y^\beta$. Show $Df(x, y) = \begin{pmatrix} \alpha x^{\alpha-1} y^\beta & x^\alpha \beta y^{\beta-1} \end{pmatrix}$.
15. Let $f(x, y) = \begin{pmatrix} x^2 \sin(y) \\ x^2 + y \end{pmatrix}$. Find $Df(x, y)$.
16. Let $f(x, y) = \sqrt{x} \sqrt[3]{y}$. Find the approximate change in f when (x, y) goes from $(4, 8)$ to $(4.01, 7.99)$.
17. Suppose f is differentiable and g is also differentiable, g having values in \mathbb{R}^3 and f having values in \mathbb{R} . Find $D(fg)$ directly from the definition. Assume both functions are defined on an open subset of \mathbb{R}^n .
18. Show, using the above definition, that if f is differentiable, then so is $t \rightarrow f(t)^n$ for any positive integer and in fact the derivative of this function is $nf(t)^{n-1} f'(t)$.
19. Suppose f is a scalar valued function of two variables which is differentiable. Show that $(x, y) \rightarrow (f(x, y))^n$ is also differentiable and its derivative equals

$$nf(x, y)^{n-1} Df(x, y)$$

20. 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

$$\lim_{h \rightarrow 0} \frac{f(ha, hb) - f(0, 0)}{h} = 0$$

for (a, b) an arbitrary unit vector. Thus the directional derivative exists at $(0, 0)$ in every direction, but f is not even continuous there.

22.4 C^1 Functions

Most of the time, there is an easier way to conclude that a derivative exists and to find it. It involves the notion of a C^1 function.

Definition 22.4.1 When $\mathbf{f} : U \rightarrow \mathbb{R}^p$ for U an open subset of \mathbb{R}^n and the vector valued functions $\frac{\partial \mathbf{f}}{\partial x_i}$ are all continuous, (equivalently each $\frac{\partial f_i}{\partial x_j}$ is continuous), the function is said to be $C^1(U)$. If all the partial derivatives up to order k exist and are continuous, then the function is said to be C^k .

It turns out that for a C^1 function, all you have to do is write the matrix described in Theorem 22.2.4 and this will be the derivative. There is no question of existence for the derivative for such functions. This is the importance of the next theorem.

Theorem 22.4.2 Suppose $\mathbf{f} : U \rightarrow \mathbb{R}^p$ where U is an open set in \mathbb{R}^n . Suppose also that all partial derivatives of \mathbf{f} exist on U and are continuous. Then \mathbf{f} is differentiable at every point of U .

Proof: If you fix all the variables but one, you can apply the fundamental theorem of calculus as follows.

$$\mathbf{f}(\mathbf{x} + v_k \mathbf{e}_k) - \mathbf{f}(\mathbf{x}) = \int_0^1 \frac{\partial \mathbf{f}}{\partial x_k}(\mathbf{x} + tv_k \mathbf{e}_k) v_k dt. \quad (22.7)$$

Here is why. Let $\mathbf{h}(t) = \mathbf{f}(\mathbf{x} + tv_k \mathbf{e}_k)$. Then

$$\frac{\mathbf{h}(t+h) - \mathbf{h}(t)}{h} = \frac{\mathbf{f}(\mathbf{x} + tv_k \mathbf{e}_k + hv_k \mathbf{e}_k) - \mathbf{f}(\mathbf{x} + tv_k \mathbf{e}_k)}{hv_k} v_k$$

and so, taking the limit as $h \rightarrow 0$ yields

$$\mathbf{h}'(t) = \frac{\partial \mathbf{f}}{\partial x_k}(\mathbf{x} + tv_k \mathbf{e}_k) v_k$$

Therefore,

$$\mathbf{f}(\mathbf{x} + v_k \mathbf{e}_k) - \mathbf{f}(\mathbf{x}) = \mathbf{h}(1) - \mathbf{h}(0) = \int_0^1 \mathbf{h}'(t) dt = \int_0^1 \frac{\partial \mathbf{f}}{\partial x_k}(\mathbf{x} + tv_k \mathbf{e}_k) v_k dt.$$

Now I will use this observation to prove the theorem. Let $\mathbf{v} = (v_1, \dots, v_n)$ with $|\mathbf{v}|$ sufficiently small. Thus $\mathbf{v} = \sum_{k=1}^n v_k \mathbf{e}_k$. For the purposes of this argument, define

$$\sum_{k=n+1}^n v_k \mathbf{e}_k \equiv \mathbf{0}.$$

Then with this convention,

$$\begin{aligned}
 f(\mathbf{x} + \mathbf{v}) - f(\mathbf{x}) &= \sum_{i=1}^n \left(f\left(\mathbf{x} + \sum_{k=i}^n v_k \mathbf{e}_k\right) - f\left(\mathbf{x} + \sum_{k=i+1}^n v_k \mathbf{e}_k\right) \right) \\
 &= \sum_{i=1}^n \int_0^1 \frac{\partial f}{\partial x_i} \left(\mathbf{x} + \sum_{k=i+1}^n v_k \mathbf{e}_k + tv_i \mathbf{e}_i \right) v_i dt \\
 &= \sum_{i=1}^n \int_0^1 \left(\frac{\partial f}{\partial x_i} \left(\mathbf{x} + \sum_{k=i+1}^n v_k \mathbf{e}_k + tv_i \mathbf{e}_i \right) v_i - \frac{\partial f}{\partial x_i}(\mathbf{x}) v_i \right) dt \\
 &\quad + \sum_{i=1}^n \int_0^1 \frac{\partial f}{\partial x_i}(\mathbf{x}) v_i dt = \sum_{i=1}^n \frac{\partial f}{\partial x_i}(\mathbf{x}) v_i \\
 &\quad + \sum_{i=1}^n \int_0^1 \left(\frac{\partial f}{\partial x_i} \left(\mathbf{x} + \sum_{k=i+1}^n v_k \mathbf{e}_k + tv_i \mathbf{e}_i \right) - \frac{\partial f}{\partial x_i}(\mathbf{x}) \right) v_i dt = \sum_{i=1}^n \frac{\partial f}{\partial x_i}(\mathbf{x}) v_i + o(\mathbf{v})
 \end{aligned}$$

and this shows f is differentiable at \mathbf{x} .

Some explanation of the step to the last line is in order. The messy thing at the end is $o(\mathbf{v})$ because of the continuity of the partial derivatives. To see this, consider one term. By Proposition 16.2.2,

$$\begin{aligned}
 &\left| \int_0^1 \left(\frac{\partial f}{\partial x_i} \left(\mathbf{x} + \sum_{k=i+1}^n v_k \mathbf{e}_k + tv_i \mathbf{e}_i \right) - \frac{\partial f}{\partial x_i}(\mathbf{x}) \right) v_i dt \right| \\
 &\leq \sqrt{p} \int_0^1 \left| \frac{\partial f}{\partial x_i} \left(\mathbf{x} + \sum_{k=i+1}^n v_k \mathbf{e}_k + tv_i \mathbf{e}_i \right) - \frac{\partial f}{\partial x_i}(\mathbf{x}) \right| dt |\mathbf{v}|
 \end{aligned}$$

Thus, dividing by $|\mathbf{v}|$ and taking a limit as $|\mathbf{v}| \rightarrow 0$, this converges to 0 due to continuity of the partial derivatives of f . The messy term is thus a finite sum of $o(\mathbf{v})$ terms and is therefore $o(\mathbf{v})$. ■

Here is an example to illustrate.

Example 22.4.3 Let $f(x, y) = \begin{pmatrix} x^2 y + y^2 \\ y^3 x \end{pmatrix}$. Find $Df(x, y)$.

From Theorem 22.4.2 this function is differentiable because all possible partial derivatives are continuous. Thus

$$Df(x, y) = \begin{pmatrix} 2xy & x^2 + 2y \\ y^3 & 3y^2 x \end{pmatrix}.$$

In particular,

$$Df(1, 2) = \begin{pmatrix} 4 & 5 \\ 8 & 12 \end{pmatrix}.$$

Here is another example.

Example 22.4.4 Let $f(x_1, x_2, x_3) = \begin{pmatrix} x_1^2 x_2 + x_2^2 \\ x_2 x_1 + x_3 \\ \sin(x_1 x_2 x_3) \end{pmatrix}$. Find $Df(x_1, x_2, x_3)$.

All possible partial derivatives are continuous, so the function is differentiable. The matrix for this derivative is therefore the following 3×3 matrix

$$\begin{pmatrix} 2x_1x_2 & x_1^2 + 2x_2 & 0 \\ x_2 & x_1 & 1 \\ x_2x_3 \cos(x_1x_2x_3) & x_1x_3 \cos(x_1x_2x_3) & x_1x_2 \cos(x_1x_2x_3) \end{pmatrix}$$

Example 22.4.5 Suppose $f(x, y, z) = xy + z^2$. Find $Df(1, 2, 3)$.

Taking the partial derivatives of f , $f_x = y$, $f_y = x$, $f_z = 2z$. These are all continuous. Therefore, the function has a derivative and $f_x(1, 2, 3) = 1$, $f_y(1, 2, 3) = 2$, and $f_z(1, 2, 3) = 6$. Therefore, $Df(1, 2, 3)$ is given by $Df(1, 2, 3) = (1, 2, 6)$. Also, for (x, y, z) close to $(1, 2, 3)$,

$$\begin{aligned} f(x, y, z) &\approx f(1, 2, 3) + 1(x - 1) + 2(y - 2) + 6(z - 3) \\ &= 11 + 1(x - 1) + 2(y - 2) + 6(z - 3) = -12 + x + 2y + 6z \end{aligned}$$

When a function is differentiable at \mathbf{x}_0 , it follows the function must be continuous there. This is the content of the following important lemma.

Lemma 22.4.6 Let $\mathbf{f} : U \rightarrow \mathbb{R}^q$ where U is an open subset of \mathbb{R}^p . If \mathbf{f} is differentiable at \mathbf{x} , then \mathbf{f} is continuous at \mathbf{x} . In fact, there is a constant C such that if $|\mathbf{v}|$ is sufficiently small, then $|\mathbf{f}(\mathbf{x} + \mathbf{v}) - \mathbf{f}(\mathbf{x})| \leq C|\mathbf{v}|$.

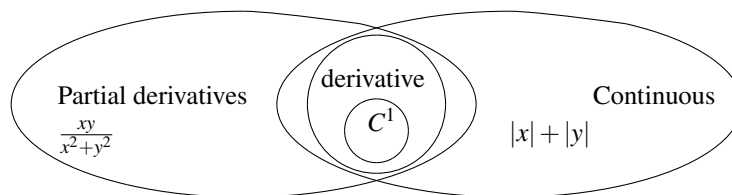
Proof: From the definition of what it means to be differentiable and Lemma 22.2.2, if $|\mathbf{v}|$ is small enough,

$$|\mathbf{f}(\mathbf{x} + \mathbf{v}) - \mathbf{f}(\mathbf{x})| = |D\mathbf{f}(\mathbf{x})(\mathbf{v}) + \mathbf{o}(\mathbf{v})| \leq |D\mathbf{f}(\mathbf{x})(\mathbf{v})| + |\mathbf{v}| \leq \tilde{C}|\mathbf{v}| + |\mathbf{v}| \equiv C|\mathbf{v}| \quad \blacksquare$$

Note that this also says that if $|\mathbf{v}|$ is small enough, then

$$\frac{|\mathbf{f}(\mathbf{x} + \mathbf{v}) - \mathbf{f}(\mathbf{x})|}{|\mathbf{v}|} \leq C \quad (22.8)$$

There have been quite a few terms defined. First there was the concept of continuity. Next the concept of partial or directional derivative. Next there was the concept of differentiability and the derivative being a linear transformation determined by a certain matrix. Finally, it was shown that if a function is C^1 , then it has a derivative. To give a rough idea of the relationships of these topics, here is a picture.



You might ask whether there are examples of functions which are differentiable but not C^1 . Of course there are. In fact, Example 22.2.9 is just such an example as explained earlier. Then you should verify that $f'(x)$ exists for all $x \in \mathbb{R}$ but f' fails to be continuous at $x = 0$. Thus the function is differentiable at every point of \mathbb{R} but fails to be C^1 because the derivative is not continuous at 0.

Example 22.4.7 Find an example of a function which is not differentiable at $(0,0)$ even though both partial derivatives exist at this point and the function is continuous at this point.

Here is a simple example.

$$f(x, y) \equiv \begin{cases} x \sin\left(\frac{1}{xy}\right) & \text{if } xy \neq 0 \\ 0 & \text{if } xy = 0 \end{cases}$$

To see this works, note that f is defined everywhere and $|f(x, y)| \leq |x|$ so clearly f is continuous at $(0, 0)$.

$$\frac{f(x, 0) - f(0, 0)}{x} = \frac{0 - 0}{x} = 0, \quad \frac{f(0, y) - f(0, 0)}{y} = \frac{0 - 0}{y} = 0$$

and so $f_x(0, 0) = 0$ and $f_y(0, 0) = 0$. Thus the partial derivatives exist. However, the function is not differentiable at $(0, 0)$ because

$$\lim_{(x, y) \rightarrow (0, 0)} \frac{x \sin\left(\frac{1}{xy}\right)}{|(x, y)|}$$

does not even exist, much less equals 0. To see this, let $x = y$ and let $x \rightarrow 0$.

22.5 The Chain Rule

22.5.1 The Chain Rule for Functions of One Variable

First recall the chain rule for a function of one variable. Consider the following picture.

$$I \xrightarrow{g} J \xrightarrow{f} \mathbb{R}$$

Here I and J are open intervals and it is assumed that $g(I) \subseteq J$. The chain rule says that if $f'(g(x))$ exists and $g'(x)$ exists for $x \in I$, then the composition, $f \circ g$ also has a derivative at x and

$$(f \circ g)'(x) = f'(g(x)) g'(x).$$

Recall that $f \circ g$ is the name of the function defined by $f \circ g(x) \equiv f(g(x))$. In the notation of this chapter, the chain rule is written as

$$Df(g(x))Dg(x) = D(f \circ g)(x). \quad (22.9)$$

22.5.2 The Chain Rule for Functions of Many Variables

Let $U \subseteq \mathbb{R}^n$ and $V \subseteq \mathbb{R}^p$ be open sets and let \mathbf{f} be a function defined on V having values in \mathbb{R}^q while \mathbf{g} is a function defined on U such that $\mathbf{g}(U) \subseteq V$ as in the following picture.

$$U \xrightarrow{\mathbf{g}} V \xrightarrow{\mathbf{f}} \mathbb{R}^q$$

The chain rule says that if the linear transformations (matrices) on the left in 22.9 both exist then the same formula holds in this more general case. Thus

$$Df(g(x))Dg(x) = D(f \circ g)(x)$$

Note this all makes sense because $Df(g(x))$ is a $q \times p$ matrix and $Dg(x)$ is a $p \times n$ matrix. Remember it is all right to do $(q \times p)(p \times n)$. The middle numbers match.

It turns out that the chain rule is an easy computation once you have the following lemma. The rough idea is as follows. Here g is differentiable at x .

$$\frac{|o(g(x+v) - g(x))|}{|v|} = \frac{\overbrace{|o(g(x+v) - g(x))|}^{\rightarrow 0 \text{ as } v \rightarrow 0}}{\overbrace{|g(x+v) - g(x)|}^{\text{bounded by 22.8}}} \frac{|g(x+v) - g(x)|}{|v|}$$

Lemma 22.5.1 *Let $g : U \rightarrow \mathbb{R}^p$ where U is an open set in \mathbb{R}^n and suppose g has a derivative at $x \in U$. Then $o(g(x+v) - g(x)) = o(v)$.*

Proof: Let

$$H(v) \equiv \begin{cases} \frac{|o(g(x+v) - g(x))|}{|g(x+v) - g(x)|} & \text{if } g(x+v) - g(x) \neq 0 \\ 0 & \text{if } g(x+v) - g(x) = 0 \end{cases}$$

Then $\lim_{v \rightarrow 0} H(v) = 0$ because of continuity of g at x and from 22.8,

$$\frac{|o(g(x+v) - g(x))|}{|v|} = H(v) \frac{|g(x+v) - g(x)|}{|v|} \leq CH(v)$$

Therefore,

$$\lim_{v \rightarrow 0} \frac{|o(g(x+v) - g(x))|}{|v|} = 0. \blacksquare$$

Now with this lemma, the chain rule is as follows.

Theorem 22.5.2 (Chain rule) *Let U be an open set in \mathbb{R}^n , let V be an open set in \mathbb{R}^p , let $g : U \rightarrow \mathbb{R}^p$ be such that $g(U) \subseteq V$, and let $f : V \rightarrow \mathbb{R}^q$. Suppose $Dg(x)$ exists for some $x \in U$ and that $Df(g(x))$ exists. Then $D(f \circ g)(x)$ exists and furthermore,*

$$D(f \circ g)(x) = Df(g(x))Dg(x). \quad (22.10)$$

In particular,

$$\frac{\partial (f \circ g)(x)}{\partial x_j} = \sum_{i=1}^p \frac{\partial f(g(x))}{\partial y_i} \frac{\partial g_i(x)}{\partial x_j}. \quad (22.11)$$

Proof: From the assumption that $Df(g(x))$ exists,

$$\begin{aligned} f(g(x+v)) &= f(g(x)) + Df(g(x))(g(x+v) - g(x)) + o(g(x+v) - g(x)) \\ &= f(g(x)) + Df(g(x))(Dg(x)v + o(v)) + o(g(x+v) - g(x)) \end{aligned}$$

which by Lemma 22.5.1 equals

$$\begin{aligned} &= f(g(x)) + Df(g(x))Dg(x)v + Df(g(x))o(v) + o(v) \\ &= f(g(x)) + Df(g(x))Dg(x)v + o(v) \end{aligned}$$

and this shows

$$D(f \circ g)(x) = Df(g(x))Dg(x)$$

from the definition of the derivative and its uniqueness established in Theorem 22.2.4 on Page 481. ■

There is an easy way to remember this in terms of the repeated index summation convention presented earlier. Let $y = g(x)$ and $z = f(y)$. Then the above says

$$\frac{\partial z}{\partial y_i} \frac{\partial y_i}{\partial x_k} = \frac{\partial z}{\partial x_k}. \quad (22.12)$$

Remember there is a sum on the repeated index. In particular, for each index r ,

$$\frac{\partial z_r}{\partial y_i} \frac{\partial y_i}{\partial x_k} = \frac{\partial z_r}{\partial x_k}.$$

The proof of this major theorem will be given later. It will include the chain rule for functions of one variable as a special case. First here are some examples.

Example 22.5.3 Let $f(u, v) = \sin(uv)$ and let $u(x, y, t) = t \sin x + \cos y$ and $v(x, y, t, s) = s \tan x + y^2 + ts$. Letting $z = f(u, v)$ where u, v are as just described, find $\frac{\partial z}{\partial t}$ and $\frac{\partial z}{\partial x}$.

From 22.12, $\frac{\partial z}{\partial t} = \frac{\partial z}{\partial u} \frac{\partial u}{\partial t} + \frac{\partial z}{\partial v} \frac{\partial v}{\partial t} = v \cos(uv) \sin(x) + us \cos(uv)$. Here $y_1 = u, y_2 = v, t = x_k$. Also,

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial u} \frac{\partial u}{\partial x} + \frac{\partial z}{\partial v} \frac{\partial v}{\partial x} = v \cos(uv) t \cos(x) + us \sec^2(x) \cos(uv).$$

Clearly you can continue in this way, taking partial derivatives with respect to any of the other variables.

Example 22.5.4 Let $w = f(u_1, u_2) = u_2 \sin(u_1)$ and $u_1 = x^2 y + z, u_2 = \sin(xy)$. Find $\frac{\partial w}{\partial x}$, $\frac{\partial w}{\partial y}$, and $\frac{\partial w}{\partial z}$.

The derivative of f is of the form (w_x, w_y, w_z) and so it suffices to find the derivative of f using the chain rule. You need to find $Df(u_1, u_2)Dg(x, y, z)$ where

$$g(x, y) = \begin{pmatrix} x^2 y + z \\ \sin(xy) \end{pmatrix}.$$

Then

$$Dg(x, y, z) = \begin{pmatrix} 2xy & x^2 & 1 \\ y \cos(xy) & x \cos(xy) & 0 \end{pmatrix}.$$

Also $Df(u_1, u_2) = (u_2 \cos(u_1), \sin(u_1))$. Therefore, the derivative is

$$\begin{aligned} Df(u_1, u_2)Dg(x, y, z) &= (u_2 \cos(u_1), \sin(u_1)) \begin{pmatrix} 2xy & x^2 & 1 \\ y \cos(xy) & x \cos(xy) & 0 \end{pmatrix} \\ &= (2u_2(\cos u_1)xy + (\sin u_1)y \cos xy, u_2(\cos u_1)x^2 + (\sin u_1)x \cos xy, u_2 \cos u_1) \\ &= (w_x, w_y, w_z) \end{aligned}$$

Thus

$$\begin{aligned}\frac{\partial w}{\partial x} &= 2u_2(\cos u_1)xy + (\sin u_1)y\cos xy = 2(\sin(xy))(\cos(x^2y+z))xy \\ &\quad + (\sin(x^2y+z))y\cos xy.\end{aligned}$$

Similarly, you can find the other partial derivatives of w in terms of substituting in for u_1 and u_2 in the above. Note

$$\frac{\partial w}{\partial x} = \frac{\partial w}{\partial u_1} \frac{\partial u_1}{\partial x} + \frac{\partial w}{\partial u_2} \frac{\partial u_2}{\partial x}.$$

In fact, in general if you have

$$w = f(u_1, u_2)$$

and $g(x, y, z) = \begin{pmatrix} u_1(x, y, z) \\ u_2(x, y, z) \end{pmatrix}$, then $D(f \circ g)(x, y, z)$ is of the form

$$\begin{aligned}&\begin{pmatrix} w_{u_1} & w_{u_2} \end{pmatrix} \begin{pmatrix} u_{1x} & u_{1y} & u_{1z} \\ u_{2x} & u_{2y} & u_{2z} \end{pmatrix} \\ &= \begin{pmatrix} w_{u_1}u_{1x} + w_{u_2}u_{2x} & w_{u_1}u_{1y} + w_{u_2}u_{2y} & w_{u_1}u_{1z} + w_{u_2}u_{2z} \end{pmatrix}.\end{aligned}$$

Example 22.5.5 Let $w = f(u_1, u_2, u_3) = u_1^2 + u_3 + u_2$ and

$$g(x, y, z) = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} x + 2yz \\ x^2 + y \\ z^2 + x \end{pmatrix}$$

Find $\frac{\partial w}{\partial x}$ and $\frac{\partial w}{\partial z}$

By the chain rule,

$$\begin{aligned}(w_x, w_y, w_z) &= \begin{pmatrix} w_{u_1} & w_{u_2} & w_{u_3} \end{pmatrix} \begin{pmatrix} u_{1x} & u_{1y} & u_{1z} \\ u_{2x} & u_{2y} & u_{2z} \\ u_{3x} & u_{3y} & u_{3z} \end{pmatrix} = \\ &\begin{pmatrix} w_{u_1}u_{1x} + w_{u_2}u_{2x} + w_{u_3}u_{3x}, w_{u_1}u_{1y} + w_{u_2}u_{2y} + w_{u_3}u_{3y}, \\ w_{u_1}u_{1z} + w_{u_2}u_{2z} + w_{u_3}u_{3z} \end{pmatrix}\end{aligned}$$

Note the pattern,

$$\begin{aligned}w_x &= w_{u_1}u_{1x} + w_{u_2}u_{2x} + w_{u_3}u_{3x}, \\ w_y &= w_{u_1}u_{1y} + w_{u_2}u_{2y} + w_{u_3}u_{3y}, \\ w_z &= w_{u_1}u_{1z} + w_{u_2}u_{2z} + w_{u_3}u_{3z}.\end{aligned}$$

Therefore,

$$w_x = 2u_1(1) + 1(2x) + 1(1) = 2(x + 2yz) + 2x + 1 = 4x + 4yz + 1$$

and

$$w_z = 2u_1(2y) + 1(0) + 1(2z) = 4(x + 2yz)y + 2z = 4yx + 8y^2z + 2z.$$

Of course to find all the partial derivatives at once, you just use the chain rule. Thus you would get

$$\begin{aligned} \begin{pmatrix} w_x & w_y & w_z \end{pmatrix} &= \begin{pmatrix} 2u_1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2z & 2y \\ 2x & 1 & 0 \\ 1 & 0 & 2z \end{pmatrix} \\ &= \begin{pmatrix} 2u_1 + 2x + 1 & 4u_1z + 1 & 4u_1y + 2z \end{pmatrix} \\ &= \begin{pmatrix} 4x + 4yz + 1 & 4zx + 8yz^2 + 1 & 4yx + 8y^2z + 2z \end{pmatrix} \end{aligned}$$

Example 22.5.6 Let $\mathbf{f}(u_1, u_2) = \begin{pmatrix} u_1^2 + u_2 \\ \sin(u_2) + u_1 \end{pmatrix}$ and

$$\mathbf{g}(x_1, x_2, x_3) = \begin{pmatrix} u_1(x_1, x_2, x_3) \\ u_2(x_1, x_2, x_3) \end{pmatrix} = \begin{pmatrix} x_1x_2 + x_3 \\ x_2^2 + x_1 \end{pmatrix}.$$

Find $D(\mathbf{f} \circ \mathbf{g})(x_1, x_2, x_3)$.

To do this,

$$D\mathbf{f}(u_1, u_2) = \begin{pmatrix} 2u_1 & 1 \\ 1 & \cos u_2 \end{pmatrix}, D\mathbf{g}(x_1, x_2, x_3) = \begin{pmatrix} x_2 & x_1 & 1 \\ 1 & 2x_2 & 0 \end{pmatrix}.$$

Then

$$D\mathbf{f}(\mathbf{g}(x_1, x_2, x_3)) = \begin{pmatrix} 2(x_1x_2 + x_3) & 1 \\ 1 & \cos(x_2^2 + x_1) \end{pmatrix}$$

and so by the chain rule,

$$\begin{aligned} D(\mathbf{f} \circ \mathbf{g})(x_1, x_2, x_3) &= \overbrace{\begin{pmatrix} 2(x_1x_2 + x_3) & 1 \\ 1 & \cos(x_2^2 + x_1) \end{pmatrix}}^{D\mathbf{f}(\mathbf{g}(\mathbf{x}))} \overbrace{\begin{pmatrix} x_2 & x_1 & 1 \\ 1 & 2x_2 & 0 \end{pmatrix}}^{D\mathbf{g}(\mathbf{x})} \\ &= \begin{pmatrix} (2x_1x_2 + 2x_3)x_2 + 1 & (2x_1x_2 + 2x_3)x_1 + 2x_2 & 2x_1x_2 + 2x_3 \\ x_2 + \cos(x_2^2 + x_1) & x_1 + 2x_2(\cos(x_2^2 + x_1)) & 1 \end{pmatrix} \end{aligned}$$

Therefore, in particular,

$$\frac{\partial f_1 \circ \mathbf{g}}{\partial x_1}(x_1, x_2, x_3) = (2x_1x_2 + 2x_3)x_2 + 1,$$

$$\frac{\partial f_2 \circ \mathbf{g}}{\partial x_3}(x_1, x_2, x_3) = 1, \frac{\partial f_2 \circ \mathbf{g}}{\partial x_2}(x_1, x_2, x_3) = x_1 + 2x_2(\cos(x_2^2 + x_1)).$$

etc.

In different notation, let $\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \mathbf{f}(u_1, u_2) = \begin{pmatrix} u_1^2 + u_2 \\ \sin(u_2) + u_1 \end{pmatrix}$. Then

$$\frac{\partial z_1}{\partial x_1} = \frac{\partial z_1}{\partial u_1} \frac{\partial u_1}{\partial x_1} + \frac{\partial z_1}{\partial u_2} \frac{\partial u_2}{\partial x_1} = 2u_1x_2 + 1 = 2(x_1x_2 + x_3)x_2 + 1.$$

Example 22.5.7 Let

$$\mathbf{f}(u_1, u_2, u_3) = \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \begin{pmatrix} u_1^2 + u_2 u_3 \\ u_1^2 + u_2^3 \\ \ln(1 + u_3^2) \end{pmatrix}$$

and let

$$\mathbf{g}(x_1, x_2, x_3, x_4) = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} x_1 + x_2^2 + \sin(x_3) + \cos(x_4) \\ x_4^2 - x_1 \\ x_3^2 + x_4 \end{pmatrix}.$$

Find $(\mathbf{f} \circ \mathbf{g})'(x)$.

$$D\mathbf{f}(\mathbf{u}) = \begin{pmatrix} 2u_1 & u_3 & u_2 \\ 2u_1 & 3u_2^2 & 0 \\ 0 & 0 & \frac{2u_3}{(1+u_3^2)} \end{pmatrix}$$

Similarly,

$$D\mathbf{g}(\mathbf{x}) = \begin{pmatrix} 1 & 2x_2 & \cos(x_3) & -\sin(x_4) \\ -1 & 0 & 0 & 2x_4 \\ 0 & 0 & 2x_3 & 1 \end{pmatrix}.$$

Then by the chain rule, $D(\mathbf{f} \circ \mathbf{g})(x) = D\mathbf{f}(\mathbf{u})D\mathbf{g}(x)$ where $\mathbf{u} = \mathbf{g}(x)$ as described above.

Thus $D(\mathbf{f} \circ \mathbf{g})(x) =$

$$\begin{pmatrix} 2u_1 & u_3 & u_2 \\ 2u_1 & 3u_2^2 & 0 \\ 0 & 0 & \frac{2u_3}{(1+u_3^2)} \end{pmatrix} \begin{pmatrix} 1 & 2x_2 & \cos(x_3) & -\sin(x_4) \\ -1 & 0 & 0 & 2x_4 \\ 0 & 0 & 2x_3 & 1 \end{pmatrix} \\ = \begin{pmatrix} 2u_1 - u_3 & 4u_1x_2 & 2u_1 \cos x_3 + 2u_2x_3 & -2u_1 \sin x_4 + 2u_3x_4 + u_2 \\ 2u_1 - 3u_2^2 & 4u_1x_2 & 2u_1 \cos x_3 & -2u_1 \sin x_4 + 6u_2^2x_4 \\ 0 & 0 & 4\frac{u_3}{1+u_3^2}x_3 & 2\frac{u_3}{1+u_3^2} \end{pmatrix} \quad (22.13)$$

where each u_i is given by the above formulas. Thus $\frac{\partial z_1}{\partial x_1}$ equals

$$\begin{aligned} 2u_1 - u_3 &= 2(x_1 + x_2^2 + \sin(x_3) + \cos(x_4)) - (x_3^2 + x_4) \\ &= 2x_1 + 2x_2^2 + 2\sin x_3 + 2\cos x_4 - x_3^2 - x_4. \end{aligned}$$

while $\frac{\partial z_2}{\partial x_4}$ equals

$$-2u_1 \sin x_4 + 6u_2^2 x_4 = -2(x_1 + x_2^2 + \sin(x_3) + \cos(x_4)) \sin(x_4) + 6(x_4^2 - x_1)^2 x_4.$$

If you wanted $\frac{\partial z}{\partial x_2}$ it would be the second column of the above matrix in 22.13. Thus $\frac{\partial z}{\partial x_2}$ equals

$$\begin{pmatrix} \frac{\partial z_1}{\partial x_2} \\ \frac{\partial z_2}{\partial x_2} \\ \frac{\partial z_3}{\partial x_2} \end{pmatrix} = \begin{pmatrix} 4u_1x_2 \\ 4u_1x_2 \\ 0 \end{pmatrix} = \begin{pmatrix} 4(x_1 + x_2^2 + \sin(x_3) + \cos(x_4))x_2 \\ 4(x_1 + x_2^2 + \sin(x_3) + \cos(x_4))x_2 \\ 0 \end{pmatrix}$$

I hope that by now it is clear that all the information you could desire about various partial derivatives is available and it all reduces to matrix multiplication and the consideration of entries of the matrix obtained by multiplying the two derivatives.

22.6 Exercises

1. Let $z = f(x_1, \dots, x_n)$ be as given and let $x_i = g_i(t_1, \dots, t_m)$ as given. Find $\frac{\partial z}{\partial t_i}$ which is indicated.

(a) $z = x_1^3 + x_2$, $x_1 = \sin(t_1) + \cos(t_2)$, $x_2 = t_1 t_2^2$. Find $\frac{\partial z}{\partial t_1}$.

(b) $z = x_1 x_2^2$, $x_1 = t_1 t_2^2 t_3$, $x_2 = t_1 t_2^2$. Find $\frac{\partial z}{\partial t_1}$.

(c) $z = x_1 x_2^2$, $x_1 = t_1 t_2^2 t_3$, $x_2 = t_1 t_2^2$. Find $\frac{\partial z}{\partial t_1}$.

(d) $z = x_1 x_2^2$, $x_1 = t_1 t_2^2 t_3$, $x_2 = t_1 t_2^2$. Find $\frac{\partial z}{\partial t_3}$.

(e) $z = x_1^2 x_2^2$, $x_1 = t_1 t_2^2 t_3$, $x_2 = t_1 t_2^2$. Find $\frac{\partial z}{\partial t_2}$.

(f) $z = x_1^2 x_2 + x_3^2$, $x_1 = t_1 t_2$, $x_2 = t_1 t_2 t_4$, $x_3 = \sin(t_3)$. Find $\frac{\partial z}{\partial t_2}$.

(g) $z = x_1^2 x_2 + x_3^2$, $x_1 = t_1 t_2$, $x_2 = t_1 t_2 t_4$, $x_3 = \sin(t_3)$. Find $\frac{\partial z}{\partial t_3}$.

(h) $z = x_1^2 x_2 + x_3^2$, $x_1 = t_1 t_2$, $x_2 = t_1 t_2 t_4$, $x_3 = \sin(t_3)$. Find $\frac{\partial z}{\partial t_1}$.

2. Let $z = f(\mathbf{y}) = (y_1^2 + \sin y_2 + \tan y_3)$ and

$$\mathbf{y} = \mathbf{g}(\mathbf{x}) \equiv \begin{pmatrix} x_1 + x_2 \\ x_2^2 - x_1 + x_2 \\ x_2^2 + x_1 + \sin x_2 \end{pmatrix}.$$

Find $D(f \circ \mathbf{g})(\mathbf{x})$. Use to write $\frac{\partial z}{\partial x_i}$ for $i = 1, 2$.

3. Let $z = f(\mathbf{y}) = (y_1^2 + \cot y_2 + \sin y_3)$ and $\mathbf{y} = \mathbf{g}(\mathbf{x}) \equiv \begin{pmatrix} x_1 + x_4 + x_3 \\ x_2^2 - x_1 + x_2 \\ x_2^2 + x_1 + \sin x_4 \end{pmatrix}$. Find

$D(f \circ \mathbf{g})(\mathbf{x})$. Use to write $\frac{\partial z}{\partial x_i}$ for $i = 1, 2, 3, 4$.

4. Let $z = f(\mathbf{y}) = (y_1^2 + y_2^2 + \sin y_3 + y_4)$, $\mathbf{y} = \mathbf{g}(\mathbf{x}) \equiv \begin{pmatrix} x_1 + x_4 + x_3 \\ x_2^2 - x_1 + x_2 \\ x_2^2 + x_1 + \sin x_4 \\ x_4 + x_2 \end{pmatrix}$. Find the

derivative of the composition $D(f \circ \mathbf{g})(\mathbf{x})$. Use to write $\frac{\partial z}{\partial x_i}$ for $i = 1, 2, 3, 4$.

5. Let

$$\mathbf{z} = \mathbf{f}(\mathbf{y}) = \begin{pmatrix} y_1^2 + \sin y_2 + \tan y_3 \\ y_1^2 y_2 + y_3 \end{pmatrix}$$

and $\mathbf{y} = \mathbf{g}(\mathbf{x}) \equiv \begin{pmatrix} x_1 + x_2 \\ x_2^2 - x_1 + x_2 \\ x_2^2 + x_1 + \sin x_2 \end{pmatrix}$. Find $D(\mathbf{f} \circ \mathbf{g})(\mathbf{x})$. Use to write $\frac{\partial z_k}{\partial x_i}$ for

$i = 1, 2$ and $k = 1, 2$. Recall this will be of the form $\begin{pmatrix} z_{1x_1} & z_{1x_2} & z_{1x_3} \\ z_{2x_1} & z_{2x_2} & z_{2x_3} \end{pmatrix}$.

6. Let $z = \mathbf{f}(\mathbf{y}) = \begin{pmatrix} y_1^2 + \sin y_2 + \tan y_3 \\ y_1^2 y_2 + y_3 \\ \cos(y_1^2) + y_2^3 y_3 \end{pmatrix}$ and

$$\mathbf{y} = \mathbf{g}(\mathbf{x}) \equiv \begin{pmatrix} x_1 + x_4 \\ x_2^2 - x_1 + x_3 \\ x_3^2 + x_1 + \sin x_2 \end{pmatrix}.$$

Find $D(\mathbf{f} \circ \mathbf{g})(\mathbf{x})$. Use to write $\frac{\partial z_k}{\partial x_i}$ for $i = 1, 2, 3, 4$ and $k = 1, 2, 3$.

7. Give a version of the chain rule which involves three functions $\mathbf{f}, \mathbf{g}, \mathbf{h}$.

8. If $\mathbf{f} : U \rightarrow V$ and $\mathbf{f}^{-1} : V \rightarrow U$ for U, V open sets such that $\mathbf{f}, \mathbf{f}^{-1}$ are both differentiable, show that

$$\det(D\mathbf{f}(\mathbf{f}^{-1}(\mathbf{y}))) \det(D\mathbf{f}^{-1}(\mathbf{y})) = 1$$

22.6.1 Related Rates Problems

Sometimes several variables are related and, given information about how one variable is changing, you want to find how the others are changing.

Example 22.6.1 *Bernoulli's law states that in an incompressible fluid,*

$$\frac{v^2}{2g} + z + \frac{P}{\gamma} = C$$

In Bernoulli's law above, each of v, z , and P are functions of (x, y, z) , the position of a point in the fluid. Find a formula for $\frac{\partial P}{\partial x}$ in terms of the partial derivatives of the other variables.

This is an example of the chain rule. Differentiate both sides with respect to x .

$$\frac{v}{g} v_x + z_x + \frac{1}{\gamma} P_x = 0$$

and so

$$P_x = - \left(\frac{v v_x + z_x g}{g} \right) \gamma$$

Example 22.6.2 *Suppose a level curve is of the form $f(x, y) = C$ and that near a point on this level curve y is a differentiable function of x . Find $\frac{dy}{dx}$.*

This is an example of the chain rule. Differentiate both sides with respect to x . This gives

$$f_x + f_y \frac{dy}{dx} = 0.$$

Solving for $\frac{dy}{dx}$ gives

$$\frac{dy}{dx} = \frac{-f_x(x, y)}{f_y(x, y)}.$$

Example 22.6.3 *Suppose a level surface is of the form $f(x, y, z) = C$. and that near a point (x, y, z) on this level surface, z is a C^1 function of x and y . Find a formula for z_x .*

This is an example of the use of the chain rule. Differentiate both sides of the equation with respect to x . Since $y_x = 0$, $f_x + f_z z_x = 0$. Then solving for z_x ,

$$z_x = \frac{-f_x(x, y, z)}{f_z(x, y, z)}$$

Example 22.6.4 Polar coordinates are

$$x = r \cos \theta, y = r \sin \theta. \quad (22.14)$$

Thus if f is a C^1 scalar valued function you could ask to express f_x in terms of the variables r and θ . Do so.

This is an example of the chain rule. Abusing notation slightly, regard f as a function of position in the plane. This position can be described with any set of coordinates. Thus $f(x, y) = f(r, \theta)$ and so

$$f_x = f_r r_x + f_\theta \theta_x.$$

This will be done if you can find r_x and θ_x . However you must find these in terms of r and θ , not in terms of x and y . Using the chain rule on the two equations for the transformation in 22.14,

$$1 = r_x \cos \theta - (r \sin \theta) \theta_x, \quad 0 = r_x \sin \theta + (r \cos \theta) \theta_x$$

Solving these using Cramer's rule,

$$r_x = \cos(\theta), \quad \theta_x = \frac{-\sin(\theta)}{r}$$

Hence f_x in polar coordinates is

$$f_x = f_r(r, \theta) \cos(\theta) - f_\theta(r, \theta) \left(\frac{\sin(\theta)}{r} \right)$$

22.6.2 The Derivative of the Inverse Function

Example 22.6.5 Let $f : U \rightarrow V$ where U and V are open sets in \mathbb{R}^n and f is one to one and onto. Suppose also that f and f^{-1} are both differentiable. How are Df^{-1} and Df related?

This can be done as follows. From the assumptions, $x = f^{-1}(f(x))$. Let $Ix = x$. Then by Example 22.2.7 on Page 482 $DI = I$. By the chain rule,

$$I = DI = Df^{-1}(f(x))(Df(x)), \quad I = DI = Df(f^{-1}(y))Df^{-1}(y)$$

Letting $y = f(x)$, the second yields

$$I = Df(x)Df^{-1}(f(x)).$$

Therefore,

$$Df(x)^{-1} = Df^{-1}(f(x)).$$

This is equivalent to

$$Df(f^{-1}(y))^{-1} = Df^{-1}(y)$$

or

$$D\mathbf{f}(\mathbf{x})^{-1} = D\mathbf{f}^{-1}(\mathbf{y}), \mathbf{y} = \mathbf{f}(\mathbf{x}).$$

This is just like a similar situation for functions of one variable. Remember

$$(f^{-1})'(f(x)) = 1/f'(x).$$

Suppose $\mathbf{y} = \mathbf{f}(\mathbf{x})$ so that $\mathbf{x} = \mathbf{f}^{-1}(\mathbf{y})$. Then the using the repeated index convention, the above can be written as

$$\delta_{ij} = \frac{\partial x_i}{\partial y_k}(\mathbf{f}(\mathbf{x})) \frac{\partial y_k}{\partial x_j}(\mathbf{x}).$$

22.7 Exercises

1. Suppose $\mathbf{f} : U \rightarrow \mathbb{R}^q$ and let $\mathbf{x} \in U$ and \mathbf{v} be a unit vector. Show that $D_{\mathbf{v}}\mathbf{f}(\mathbf{x}) = D\mathbf{f}(\mathbf{x})\mathbf{v}$. Recall that

$$D_{\mathbf{v}}\mathbf{f}(\mathbf{x}) \equiv \lim_{t \rightarrow 0} \frac{\mathbf{f}(\mathbf{x} + t\mathbf{v}) - \mathbf{f}(\mathbf{x})}{t}.$$

2. Let $f(x, y) = \begin{cases} xy \sin(\frac{1}{x}) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$. Find where f is differentiable and compute the derivative at all these points.

3. Let

$$f(x, y) = \begin{cases} x & \text{if } |y| > |x| \\ -x & \text{if } |y| \leq |x| \end{cases}.$$

Show that f is continuous at $(0, 0)$ and that the partial derivatives exist at $(0, 0)$ but the function is not differentiable at $(0, 0)$.

4. Let

$$\mathbf{f}(x, y, z) = \begin{pmatrix} x^2 \sin y + z^3 \\ \sin(x + y) + z^3 \cos x \end{pmatrix}.$$

Find $D\mathbf{f}(1, 2, 3)$.

5. Let

$$\mathbf{f}(x, y, z) = \begin{pmatrix} x \tan y + z^3 \\ \cos(x + y) + z^3 \cos x \end{pmatrix}.$$

Find $D\mathbf{f}(x, y, z)$.

6. Let

$$\mathbf{f}(x, y, z) = \begin{pmatrix} x \sin y + z^3 \\ \sin(x + y) + z^3 \cos x \\ x^5 + y^2 \end{pmatrix}.$$

Find $D\mathbf{f}(x, y, z)$.

7. 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 all directional derivatives of f exist at $(0, 0)$, and are all equal to zero but the function is not even continuous at $(0, 0)$. Therefore, it is not differentiable. Why?

8. In the example of Problem 7 show that the partial derivatives exist but are not continuous.
9. A certain building is shaped like the top half of the ellipsoid, $\frac{x^2}{900} + \frac{y^2}{900} + \frac{z^2}{400} = 1$ determined by letting $z \geq 0$. Here dimensions are measured in feet. The building needs to be painted. The paint, when applied is about .005 feet thick. About how many cubic feet of paint will be needed. **Hint:** This is going to replace the numbers, 900 and 400 with slightly larger numbers when the ellipsoid is fattened slightly by the paint. The volume of the top half of the ellipsoid, $x^2/a^2 + y^2/b^2 + z^2/c^2 \leq 1, z \geq 0$ is $(2/3)\pi abc$.
10. Suppose $\mathbf{r}_1(t) = (\cos t, \sin t, t)$, $\mathbf{r}_2(t) = (t, 2t, 1)$, and $\mathbf{r}_3(t) = (1, t, 1)$. Find the rate of change with respect to t of the volume of the parallelepiped determined by these three vectors when $t = 1$.
11. A trash compactor is compacting a rectangular block of trash. The width is changing at the rate of -1 inches per second, the length is changing at the rate of -2 inches per second and the height is changing at the rate of -3 inches per second. How fast is the volume changing when the length is 20, the height is 10, and the width is 10?
12. A trash compactor is compacting a rectangular block of trash. The width is changing at the rate of -2 inches per second, the length is changing at the rate of -1 inches per second and the height is changing at the rate of -4 inches per second. How fast is the surface area changing when the length is 20, the height is 10, and the width is 10?
13. The ideal gas law is $PV = kT$ where k is a constant which depends on the number of moles and on the gas being considered. If V is changing at the rate of 2 cubic cm. per second and T is changing at the rate of 3 degrees Kelvin per second, how fast is the pressure changing when $T = 300$ and V equals 400 cubic cm.?
14. Let S denote a level surface of the form $f(x_1, x_2, x_3) = C$. Show that any smooth curve in the level surface is perpendicular to the gradient.
15. Suppose \mathbf{f} is a C^1 function which maps U , an open subset of \mathbb{R}^n one to one and onto V , an open set in \mathbb{R}^m such that the inverse map, \mathbf{f}^{-1} is also C^1 . What must be true of m and n ? Why? **Hint:** Consider Example 22.6.5 on Page 497. Also you can use the fact that if A is an $m \times n$ matrix which maps \mathbb{R}^n onto \mathbb{R}^m , then $m \leq n$.
16. Finish Example 22.6.4 by finding f_y in terms of θ, r . Show that $f_y = \sin(\theta) f_r + \frac{\cos(\theta)}{r} f_\theta$.

17. *Think of ∂_x as a differential operator which takes functions and differentiates them with respect to x . Thus $\partial_x f \equiv f_x$. In the context of Example 22.6.4, which is on polar coordinates, and Problem 16, explain how

$$\begin{aligned}\partial_x &= \cos(\theta) \partial_r - \frac{\sin(\theta)}{r} \partial_\theta \\ \partial_y &= \sin(\theta) \partial_r + \frac{\cos(\theta)}{r} \partial_\theta\end{aligned}$$

The Laplacian of a function u is defined as $\Delta u = u_{xx} + u_{yy}$. Use the above observation to give a formula Δu in terms of r and θ . You should get $u_{rr} + \frac{1}{r}u_r + \frac{1}{r^2}u_{\theta\theta}$. This is the formula for the Laplacian in polar coordinates.

22.8 The Gradient

Here we review the concept of the gradient and the directional derivative and prove the formula for the directional derivative discussed earlier.

Let $f : U \rightarrow \mathbb{R}$ where U is an open subset of \mathbb{R}^n and suppose f is differentiable on U . Thus if $\mathbf{x} \in U$,

$$f(\mathbf{x} + \mathbf{v}) = f(\mathbf{x}) + \sum_{j=1}^n \frac{\partial f(\mathbf{x})}{\partial x_j} v_j + o(\mathbf{v}). \quad (22.15)$$

Now we can prove the formula for the directional derivative in terms of the gradient.

Proposition 22.8.1 *If f is differentiable at \mathbf{x} and for \mathbf{v} a unit vector*

$$D_{\mathbf{v}}f(\mathbf{x}) = \nabla f(\mathbf{x}) \cdot \mathbf{v}. \quad (22.16)$$

Proof:

$$\begin{aligned}\frac{f(\mathbf{x} + t\mathbf{v}) - f(\mathbf{x})}{t} &= \frac{1}{t} \left(f(\mathbf{x}) + \sum_{j=1}^n \frac{\partial f(\mathbf{x})}{\partial x_j} t v_j + o(t\mathbf{v}) - f(\mathbf{x}) \right) \\ &= \frac{1}{t} \left(\sum_{j=1}^n \frac{\partial f(\mathbf{x})}{\partial x_j} t v_j + o(t\mathbf{v}) \right) = \sum_{j=1}^n \frac{\partial f(\mathbf{x})}{\partial x_j} v_j + \frac{o(t\mathbf{v})}{t}\end{aligned}$$

Now $\lim_{t \rightarrow 0} \frac{o(t\mathbf{v})}{t} = 0$ and so

$$D_{\mathbf{v}}f(\mathbf{x}) = \lim_{t \rightarrow 0} \frac{f(\mathbf{x} + t\mathbf{v}) - f(\mathbf{x})}{t} = \sum_{j=1}^n \frac{\partial f(\mathbf{x})}{\partial x_j} v_j = \nabla f(\mathbf{x}) \cdot \mathbf{v}$$

as claimed. ■

Example 22.8.2 *Let $f(x, y, z) = x^2 + \sin(xy) + z$. Find $D_{\mathbf{v}}f(1, 0, 1)$ where*

$$\mathbf{v} = \left(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}} \right).$$

Note this vector which is given is already a unit vector. Therefore, from the above, it is only necessary to find $\nabla f(1, 0, 1)$ and take the dot product.

$$\nabla f(x, y, z) = (2x + (\cos xy)y, (\cos xy)x, 1).$$

Therefore, $\nabla f(1, 0, 1) = (2, 1, 1)$. Therefore, the directional derivative is

$$(2, 1, 1) \cdot \left(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}} \right) = \frac{4}{3}\sqrt{3}.$$

Because of 22.16 it is easy to find the largest possible directional derivative and the smallest possible directional derivative. That which follows is a more algebraic treatment of an earlier result with the trigonometry removed.

Proposition 22.8.3 *Let $f : U \rightarrow \mathbb{R}$ be a differentiable function and let $\mathbf{x} \in U$. Then*

$$\max \{D_{\mathbf{v}}f(\mathbf{x}) : |\mathbf{v}| = 1\} = |\nabla f(\mathbf{x})| \quad (22.17)$$

and

$$\min \{D_{\mathbf{v}}f(\mathbf{x}) : |\mathbf{v}| = 1\} = -|\nabla f(\mathbf{x})|. \quad (22.18)$$

Furthermore, the maximum in 22.17 occurs when $\mathbf{v} = \nabla f(\mathbf{x}) / |\nabla f(\mathbf{x})|$ and the minimum in 22.18 occurs when $\mathbf{v} = -\nabla f(\mathbf{x}) / |\nabla f(\mathbf{x})|$.

Proof: From 22.16 and the Cauchy Schwarz inequality,

$$|D_{\mathbf{v}}f(\mathbf{x})| \leq |\nabla f(\mathbf{x})|$$

and so for any choice of \mathbf{v} with $|\mathbf{v}| = 1$,

$$-|\nabla f(\mathbf{x})| \leq D_{\mathbf{v}}f(\mathbf{x}) \leq |\nabla f(\mathbf{x})|.$$

The proposition is proved by noting that if $\mathbf{v} = -\nabla f(\mathbf{x}) / |\nabla f(\mathbf{x})|$, then

$$D_{\mathbf{v}}f(\mathbf{x}) = \nabla f(\mathbf{x}) \cdot (-\nabla f(\mathbf{x}) / |\nabla f(\mathbf{x})|) = -|\nabla f(\mathbf{x})|^2 / |\nabla f(\mathbf{x})| = -|\nabla f(\mathbf{x})|$$

while if $\mathbf{v} = \nabla f(\mathbf{x}) / |\nabla f(\mathbf{x})|$, then

$$D_{\mathbf{v}}f(\mathbf{x}) = \nabla f(\mathbf{x}) \cdot (\nabla f(\mathbf{x}) / |\nabla f(\mathbf{x})|) = |\nabla f(\mathbf{x})|^2 / |\nabla f(\mathbf{x})| = |\nabla f(\mathbf{x})|. \blacksquare$$

For a different approach to the proposition, see Problem 7 which follows.

The conclusion of the above proposition is important in many physical models. For example, consider some material which is at various temperatures depending on location. Because it has cool places and hot places, it is expected that the heat will flow from the hot places to the cool places. Consider a small surface having a unit normal \mathbf{n} . Thus \mathbf{n} is a normal to this surface and has unit length. If it is desired to find the rate in calories per second at which heat crosses this little surface in the direction of \mathbf{n} it is defined as $\mathbf{J} \cdot \mathbf{n}A$ where A is the area of the surface and \mathbf{J} is called the heat flux. It is reasonable to suppose the rate at which heat flows across this surface will be largest when \mathbf{n} is in the direction of greatest rate of decrease of the temperature. In other words, heat flows most readily in the direction which involves the maximum rate of decrease in temperature. This expectation will be realized by taking $\mathbf{J} = -K\nabla u$ where K is a positive scalar function which can

depend on a variety of things. The above relation between the heat flux and ∇u is usually called the Fourier heat conduction law and the constant K is known as the coefficient of thermal conductivity. It is a material property, different for iron than for aluminum. In most applications, K is considered to be a constant but this is wrong. Experiments show that this scalar should depend on temperature. Nevertheless, things get very difficult if this dependence is allowed. The constant can depend on position in the material or even on time.

An identical relationship is usually postulated for the flow of a diffusing species. In this problem, something like a pollutant diffuses. It may be an insecticide in ground water for example. Like heat, it tries to move from areas of high concentration toward areas of low concentration. In this case $\mathbf{J} = -K\nabla c$ where c is the concentration of the diffusing species. When applied to diffusion, this relationship is known as Fick's law. Mathematically, it is indistinguishable from the problem of heat flow.

Note the importance of the gradient in formulating these models.

22.9 The Gradient and Tangent Planes

Let $S \equiv \{\mathbf{x} \in \mathbb{R}^p : g(\mathbf{x}) = 0\}$ be a level surface. We assume $\nabla g(\mathbf{y}) \neq 0$ for some $\mathbf{y} \in S$. Then a plane tangent to this level surface at \mathbf{y} will be of the form $\{\mathbf{x} \in \mathbb{R}^p : \mathbf{n} \cdot (\mathbf{x} - \mathbf{y}) = 0\}$. The problem is to find \mathbf{n} which is a vector which is perpendicular to every vector from \mathbf{y} to \mathbf{x} and we want this to be a real tangent plane. The way you can achieve this is to require that \mathbf{n} be perpendicular to the direction vector of every smooth curve through \mathbf{y} which lies in S . One such \mathbf{n} is obtained from $\nabla g(\mathbf{y})$. Indeed, if $t \rightarrow \mathbf{x}(t)$ is a curve through \mathbf{y} such that $\mathbf{x}(0) = \mathbf{y}$, then $g(\mathbf{x}(t)) = 0$ and so from the chain rule, $\nabla g(\mathbf{y}) \cdot \mathbf{x}'(0) = 0$. Thus a suitable choice for \mathbf{n} will be $\nabla g(\mathbf{y})$. Of course, this is a specious argument without the implicit function theorem which gives existence of such smooth curves in the level surface. See the appendix for this major theorem.

Example 22.9.1 Find the equation of the tangent plane to the level surface

$$f(x, y, z, w) = 6$$

of the function $f(x, y, z) = x^2 + 2y^2 + 3z^2 + w$ at the point $(1, 1, 1, 0)$.

First note that $(1, 1, 1, 0)$ is a point on this level surface. To find the desired plane it suffices to find the normal vector to the proposed plane. But $\nabla f(x, y, z, w) = (2x, 4y, 6z, 1)$ and so $\nabla f(1, 1, 1, 0) = (2, 4, 6, 1)$. Therefore, from this problem, the equation of the plane is $(2, 4, 6, 1) \cdot (x - 1, y - 1, z - 1, w) = 0$ or in other words, $2x - 12 + 4y + 6z + w = 0$. Note that this is a three dimensional plane because there are three free variables. Indeed, it is of the form $w = 12 - 4y - 6z - 2x$.

Example 22.9.2 The point $(\sqrt{3}, 1, 4)$ is on both the surfaces, $z = x^2 + y^2$ and $z = 8 - (x^2 + y^2)$. Find the cosine of the angle between the two tangent planes at this point.

Recall this is the same as the angle between two normal vectors. Of course there is some ambiguity here because if \mathbf{n} is a normal vector, then so is $-\mathbf{n}$ and replacing \mathbf{n} with $-\mathbf{n}$ in the formula for the cosine of the angle will change the sign. We agree to look for the acute angle and its cosine rather than the obtuse angle. The normals are $(2\sqrt{3}, 2, -1)$ and $(2\sqrt{3}, 2, 1)$. Therefore, the cosine of the angle desired is $\frac{(2\sqrt{3})^2 + 4 - 1}{17} = \frac{15}{17}$.

Example 22.9.3 The point $(1, \sqrt{3}, 4)$ is on the surface $z = x^2 + y^2$. Find the line perpendicular to the surface at this point.

All that is needed is a direction vector for this line. The surface is the level surface $x^2 + y^2 - z = 0$. The normal to this surface is given by the gradient at this point. Thus the desired line is $(1, \sqrt{3}, 4) + t(2, 2\sqrt{3}, -1)$.

22.10 Exercises

1. Find the gradient at the indicated point if $f =$

- | | |
|--|--|
| (a) $x^2y + z^3, (1, 1, 2)$ | (d) $\sin(xy) + z^3, (1, \pi, 1)$ |
| (b) $z \sin(x^2y) + 2^{x+y}, (1, 1, 0)$ | (e) $\ln(x + y^2)z$ |
| (c) $u \ln(x + y + z^2 + w), (x, y, z, w, u)$
$= (1, 1, 1, 1, 2)$ | (f) $z \ln(4 + \sin(xy)), (0, \pi, 1)$ |

2. Find the directional derivatives of f at the indicated point in the direction $\left(\frac{1}{2}, \frac{1}{2}, \frac{1}{\sqrt{2}}\right)$.

- | | |
|---|---|
| (a) $x^2y + z^3$ at $(1, 1, 1)$ | (e) $x^y + z$ at $(1, 1, 1)$. |
| (b) $z \sin(x^2y) + 2^{x+y}$ at $(1, 1, 0)$ | |
| (c) $xy + z^2 + 1$ at $(1, 2, 3)$ | (f) $\sin(\sin(x + y)) + z$ at the point
$(1, 0, 1)$. |
| (d) $\sin(xy) + z$ at $(0, 1, 1)$ | |

3. Find the directional derivatives of the given function at the indicated point in the indicated direction.

- | |
|--|
| (a) $\sin(x^2 + y) + z^2$ at $(0, \pi/2, 1)$ in direction of $(1, 1, 2)$. |
| (b) $x^{(x+y)} + \sin(zx)$ at $(1, 0, 0)$ in the direction of $(2, -1, 0)$. |
| (c) $z^{\sin(x)} + y$ at $(0, 1, 1)$ in the direction of $(1, 1, 3)$. |

4. Find the tangent plane to the indicated level surface at the indicated point.

- | |
|--|
| (a) $x^2y + z^3 = 2$ at $(1, 1, 1)$ |
| (b) $z \sin(x^2y) + 2^{x+y} = 2 \sin 1 + 4$ at $(1, 1, 2)$ |
| (c) $\cos(x) + z \sin(x + y) = 1$ at $(-\pi, \frac{3\pi}{2}, 2)$ |

5. The point $(1, 1, \sqrt{2})$ is a point on the level surface $x^2 + y^2 + z^2 = 4$. Find the line perpendicular to the surface at this point.

6. The level surfaces $x^2 + y^2 + z^2 = 4$ and $z + x^2 + y^2 = 4$ have the point $\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}, 1\right)$ in the curve formed by the intersection of these surfaces. Find a direction vector for this curve at this point. **Hint:** Recall the gradients of the two surfaces are perpendicular to the corresponding surfaces at this point. A direction vector for the desired curve should be perpendicular to both of these gradients.

7. For \mathbf{v} a unit vector, recall that $D_{\mathbf{v}}f(\mathbf{x}) = \nabla f(\mathbf{x}) \cdot \mathbf{v}$. It was shown above that the largest directional derivative is in the direction of the gradient and the smallest in the direction of $-\nabla f$. Establish the same result using the geometric description of the dot product, the one which says the dot product is the product of the lengths of the vectors times the cosine of the included angle.
8. * Suppose $f_1(x, y, z) = 0$ and $f_2(x, y, z) = 0$ are two level surfaces which intersect in a curve which has parametrization, $(x(t), y(t), z(t))$. Find a system of differential equations for $(x(t), y(t), z(t))$ where the point determined by $(x(t), y(t), z(t))$ as t varies, moves over the curve.

Chapter 23

Optimization

23.1 Local Extrema

The following definition describes what is meant by a local maximum or local minimum.

Definition 23.1.1 Suppose $f : D(f) \rightarrow \mathbb{R}$ where $D(f) \subseteq \mathbb{R}^p$. A point $\mathbf{x} \in D(f) \subseteq \mathbb{R}^p$ is called a **local minimum** if $f(\mathbf{x}) \leq f(\mathbf{y})$ for all $\mathbf{y} \in D(f)$ sufficiently close to \mathbf{x} . A point $\mathbf{x} \in D(f)$ is called a **local maximum** if $f(\mathbf{x}) \geq f(\mathbf{y})$ for all $\mathbf{y} \in D(f)$ sufficiently close to \mathbf{x} . A **local extremum** is a point of $D(f)$ which is either a local minimum or a local maximum. The plural for extremum is *extrema*. The plural for minimum is **minima** and the plural for maximum is **maxima**.

Procedure 23.1.2 To find candidates for local extrema which are interior points of $D(f)$ where f is a differentiable function, you simply identify those points where ∇f equals the zero vector.

To locate candidates for local extrema, for the function f , take ∇f and find where this vector equals 0.

Let \mathbf{v} be any vector in \mathbb{R}^p and suppose \mathbf{x} is a local maximum (minimum) for \mathbf{f} . Then consider the real valued function of one variable, $h(t) \equiv f(\mathbf{x} + t\mathbf{v})$ for small $|t|$. Since \mathbf{f} has a local maximum (minimum), it follows that h is a differentiable function of the single variable t for small t which has a local maximum (minimum) when $t = 0$. Therefore, $h'(0) = 0$.

$$\begin{aligned} h(\Delta t) - h(0) &= f(\mathbf{x} + \Delta t \mathbf{v}) - f(\mathbf{x}) \\ &= Df(\mathbf{x}) \Delta t \mathbf{v} + o(\Delta t) \end{aligned}$$

Now divide by Δt and let $\Delta t \rightarrow 0$ to obtain

$$0 = h'(0) = Df(\mathbf{x}) \mathbf{v}$$

and since \mathbf{v} is arbitrary, it follows $Df(\mathbf{x}) = 0$. However,

$$Df(\mathbf{x}) = \begin{pmatrix} f_{x_1}(\mathbf{x}) & \cdots & f_{x_p}(\mathbf{x}) \end{pmatrix}$$

and so $\nabla f(\mathbf{x}) = 0$. This proves the following theorem.

Theorem 23.1.3 Suppose U is an open set contained in $D(f)$ such that f is differentiable on U and suppose $\mathbf{x} \in U$ is a local minimum or local maximum for f . Then $\nabla f(\mathbf{x}) = \mathbf{0}$.

Definition 23.1.4 A **singular point** for f is a point \mathbf{x} where $\nabla f(\mathbf{x}) = \mathbf{0}$. This is also called a **critical point**. By analogy with the one variable case, a point where the gradient does not exist will also be called a critical point.

Example 23.1.5 Find the critical points for the function $f(x, y) \equiv xy - x - y$ for $x, y > 0$.

Note that here $D(f)$ is an open set and so every point is an interior point. Where is the gradient equal to zero? $f_x = y - 1 = 0$, $f_y = x - 1 = 0$, and so there is exactly one critical point $(1, 1)$.

Example 23.1.6 Find the volume of the smallest tetrahedron made up of the coordinate planes in the first octant and a plane which is tangent to the sphere $x^2 + y^2 + z^2 = 4$.

The normal to the sphere at a point (x_0, y_0, z_0) of the sphere is

$$\left(x_0, y_0, \sqrt{4 - x_0^2 - y_0^2} \right)$$

and so the equation of the tangent plane at this point is

$$x_0(x - x_0) + y_0(y - y_0) + \sqrt{4 - x_0^2 - y_0^2} \left(z - \sqrt{4 - x_0^2 - y_0^2} \right) = 0$$

When $x = y = 0$, $z = \frac{4}{\sqrt{4 - x_0^2 - y_0^2}}$. When $z = 0 = y$, $x = \frac{4}{x_0}$, and when $z = x = 0$, $y = \frac{4}{y_0}$.

Therefore, the function to minimize is

$$f(x, y) = \frac{1}{6} \frac{64}{xy\sqrt{(4 - x^2 - y^2)}}$$

This is because in beginning calculus it was shown that the volume of a pyramid is $1/3$ the area of the base times the height. Therefore, you simply need to find the gradient of this and set it equal to zero. Thus upon taking the partial derivatives, you need to have

$$\frac{-4 + 2x^2 + y^2}{x^2y(-4 + x^2 + y^2)\sqrt{(4 - x^2 - y^2)}} = 0,$$

and

$$\frac{-4 + x^2 + 2y^2}{xy^2(-4 + x^2 + y^2)\sqrt{(4 - x^2 - y^2)}} = 0.$$

Therefore, $x^2 + 2y^2 = 4$ and $2x^2 + y^2 = 4$. Thus $x = y$ and so $x = y = \frac{2}{\sqrt{3}}$. It follows from the equation for z that $z = \frac{2}{\sqrt{3}}$ also. How do you know this is not the largest tetrahedron?

Example 23.1.7 An open box is to contain 32 cubic feet. Find the dimensions which will result in the least surface area.

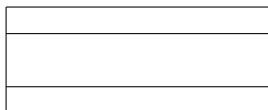
Let the height of the box be z and the length and width be x and y respectively. Then $xyz = 32$ and so $z = 32/xy$. The total area is $xy + 2xz + 2yz$ and so in terms of the two variables x and y , the area is $A = xy + \frac{64}{y} + \frac{64}{x}$. To find best dimensions you note these must result in a local minimum.

$$A_x = \frac{yx^2 - 64}{x^2} = 0, A_y = \frac{xy^2 - 64}{y^2}.$$

Therefore, $yx^2 - 64 = 0$ and $xy^2 - 64 = 0$ so $xy^2 = yx^2$. For sure the answer excludes the case where any of the variables equals zero. Therefore, $x = y$ and so $x = 4 = y$. Then $z = 2$ from the requirement that $xyz = 32$. How do you know this gives the least surface area? Why is this not the largest surface area?

23.2 Exercises

- Find the points where possible local minima or local maxima occur in the following functions.
 - $x^2 - 2x + 5 + y^2 - 4y$
 - $-xy + y^2 - y + x$
 - $3x^2 - 4xy + 2y^2 - 2y + 2x$
 - $\cos(x) + \sin(2y)$
 - $x^4 - 4x^3y + 6x^2y^2 - 4xy^3 + y^4 + x^2 - 2x$
 - $y^2x^2 - 2xy^2 + y^2$
- Find the volume of the largest box which can be inscribed in a sphere of radius a .
- Find in terms of a, b, c the volume of the largest box which can be inscribed in the ellipsoid $\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$.
- Find three numbers which add to 36 whose product is as large as possible.
- Find three numbers x, y, z such that $x^2 + y^2 + z^2 = 1$ and $x + y + z$ is as large as possible.
- Find three numbers x, y, z such that $x^2 + y^2 + z^2 = 4$ and xyz is as large as possible.
- A feeding trough in the form of a trapezoid with equal base angles is made from a long rectangular piece of metal of width 24 inches by bending up equal strips along both sides. Find the base angles and the width of these strips which will maximize the volume of the feeding trough.



- An open box (no top) is to contain 40 cubic feet. The material for the bottom costs twice as much as the material for the sides. Find the dimensions of the box which is cheapest.

9. The function $f(x, y) = 2x^2 + y^2$ is defined on the disk $x^2 + y^2 \leq 1$. Find its maximum value.
10. Find the point on the surface $z = x^2 + y + 1$ which is closest to $(0, 0, 0)$.
11. Let $L_1 = (t, 2t, 3 - t)$ and $L_2 = (2s, s + 2, 4 - s)$ be two lines. Find a pair of points, one on the first line and the other on the second such that these two points are closer together than any other pair of points on the two lines.
12. *Let

$$f(x, y) = \begin{cases} -1 & \text{if } y = x^2, x \neq 0 \\ (y - x^2)^2 & \text{if } y \neq x^2 \\ 0 & \text{if } (x, y) = (0, 0) \end{cases}$$

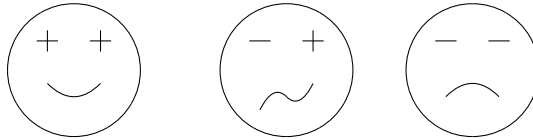
Show that $\nabla f(0, 0) = \mathbf{0}$. Now show that if (a, b) is any nonzero unit vector, the function $t \rightarrow f(ta, tb)$ has a local minimum of 0 when $t = 0$. Thus in every direction, this function has a local minimum at $(0, 0)$ but the function f does not have a local minimum at $(0, 0)$.

23.3 The Second Derivative Test

There is a version of the second derivative test in the case that the function and its first and second partial derivatives are all continuous.

Definition 23.3.1 The matrix $H(\mathbf{x})$ whose ij^{th} entry at the point \mathbf{x} is $\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x})$ is called the **Hessian matrix**. The eigenvalues of $H(\mathbf{x})$ are the solutions λ to the equation $\det(\lambda I - H(\mathbf{x})) = 0$.

The following theorem says that if all the eigenvalues of the Hessian matrix at a critical point are positive, then the critical point is a local minimum. If all the eigenvalues of the Hessian matrix at a critical point are negative, then the critical point is a local maximum. Finally, if some of the eigenvalues of the Hessian matrix at the critical point are positive and some are negative then the critical point is a saddle point. The following picture illustrates the situation.



Theorem 23.3.2 Let $f : U \rightarrow \mathbb{R}$ for U an open set in \mathbb{R}^p and let f be a C^2 function and suppose that at some $\mathbf{x} \in U$, $\nabla f(\mathbf{x}) = \mathbf{0}$. Also let μ and λ be respectively, the largest and smallest eigenvalues of the matrix $H(\mathbf{x})$. If $\lambda > 0$ then f has a local minimum at \mathbf{x} . If $\mu < 0$ then f has a local maximum at \mathbf{x} . If either λ or μ equals zero, the test fails. If $\lambda < 0$ and $\mu > 0$ there exists a direction in which when f is evaluated on the line through the critical point having this direction, the resulting function of one variable has a local minimum and there exists a direction in which when f is evaluated on the line through the critical point having this direction, the resulting function of one variable has a local maximum. This last case is called a **saddle point**.

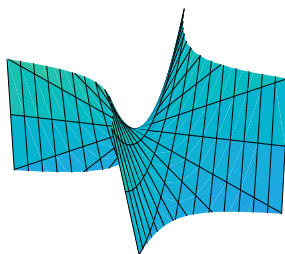
Here is an example.

Example 23.3.3 Let $f(x, y) = 10xy + y^2$. Find the critical points and determine whether they are local minima, local maxima or saddle points.

First $\nabla(10xy + y^2) = (10y, 10x + 2y)$ and so there is one critical point at the point $(0, 0)$. What is it? The Hessian matrix is

$$\begin{pmatrix} 0 & 10 \\ 10 & 2 \end{pmatrix}$$

and the eigenvalues are of different signs. Therefore, the critical point $(0, 0)$ is a saddle point. Here is a graph drawn by Matlab.



Here is another example.

Example 23.3.4 Let $f(x, y) = 2x^4 - 4x^3 + 14x^2 + 12yx^2 - 12yx - 12x + 2y^2 + 4y + 2$. Find the critical points and determine whether they are local minima, local maxima, or saddle points.

$f_x(x, y) = 8x^3 - 12x^2 + 28x + 24yx - 12y - 12$ and $f_y(x, y) = 12x^2 - 12x + 4y + 4$. The points at which both f_x and f_y equal zero are $(\frac{1}{2}, -\frac{1}{4})$, $(0, -1)$, and $(1, -1)$.

The Hessian matrix is

$$\begin{pmatrix} 24x^2 + 28 + 24y - 24x & 24x - 12 \\ 24x - 12 & 4 \end{pmatrix}$$

and the thing to determine is the sign of its eigenvalues evaluated at the critical points.

First consider the point $(\frac{1}{2}, -\frac{1}{4})$. The Hessian matrix is $\begin{pmatrix} 16 & 0 \\ 0 & 4 \end{pmatrix}$ and its eigenvalues are 16, 4 showing that this is a local minimum.

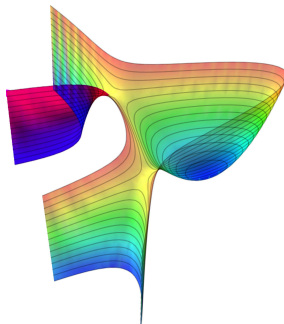
Next consider $(0, -1)$ at this point the Hessian matrix is $\begin{pmatrix} 4 & -12 \\ -12 & 4 \end{pmatrix}$ and the eigenvalues are 16, -8 . Therefore, this point is a saddle point. To determine this, find the eigenvalues.

$$\det\left(\lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 4 & -12 \\ -12 & 4 \end{pmatrix}\right) = \lambda^2 - 8\lambda - 128 = (\lambda + 8)(\lambda - 16)$$

so the eigenvalues are -8 and 16 as claimed.

Finally consider the point $(1, -1)$. At this point the Hessian is $\begin{pmatrix} 4 & 12 \\ 12 & 4 \end{pmatrix}$ and the eigenvalues are 16, -8 so this point is also a saddle point.

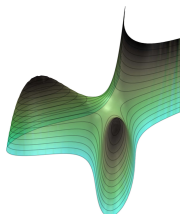
Below is a graph of this function which illustrates the behavior near saddle points.



Of course sometimes the second derivative test is inadequate to determine what is going on. This should be no surprise since this was the case even for a function of one variable. For a function of two variables, a nice example is the monkey saddle.

Example 23.3.5 Suppose $f(x, y) = 6xy^2 - 2x^3 - 3y^4$. Show that $(0, 0)$ is a critical point for which the second derivative test gives no information.

Before doing anything it might be interesting to look at the graph of this function of two variables plotted using a computer algebra system.



This picture should indicate why this is called a monkey saddle. It is because the monkey can sit in the saddle and have a place for his tail. Now to see $(0, 0)$ is a critical point, note that $f_x(0, 0) = f_y(0, 0) = 0$ because $f_x(x, y) = 6y^2 - 6x^2$, $f_y(x, y) = 12xy - 12y^3$ and so $(0, 0)$ is a critical point. So are $(1, 1)$ and $(1, -1)$. Now $f_{xx}(0, 0) = 0$ and so are $f_{xy}(0, 0)$ and $f_{yy}(0, 0)$. Therefore, the Hessian matrix is the zero matrix and clearly has only the zero eigenvalue. Therefore, the second derivative test is totally useless at this point.

However, suppose you took $x = t$ and $y = t$ and evaluated this function on this line. This reduces to $h(t) = f(t, t) = 4t^3 - 3t^4$, which is strictly increasing near $t = 0$. This shows the critical point $(0, 0)$ of f is neither a local max. nor a local min. Next let $x = 0$ and $y = t$. Then $p(t) \equiv f(0, t) = -3t^4$. Therefore, along the line, $(0, t)$, f has a local maximum at $(0, 0)$.

Example 23.3.6 Find the critical points of the following function of three variables and classify them as local minimums, local maximums or saddle points.

$$f(x, y, z) = \frac{5}{6}x^2 + 4x + 16 - \frac{7}{3}xy - 4y - \frac{4}{3}xz + 12z + \frac{5}{6}y^2 - \frac{4}{3}zy + \frac{1}{3}z^2$$

First you need to locate the critical points. This involves taking the gradient.

$$\begin{aligned} & \nabla \left(\frac{5}{6}x^2 + 4x + 16 - \frac{7}{3}xy - 4y - \frac{4}{3}xz + 12z + \frac{5}{6}y^2 - \frac{4}{3}zy + \frac{1}{3}z^2 \right) \\ &= \left(\frac{5}{3}x + 4 - \frac{7}{3}y - \frac{4}{3}z, -\frac{7}{3}x - 4 + \frac{5}{3}y - \frac{4}{3}z, -\frac{4}{3}x + 12 - \frac{4}{3}y + \frac{2}{3}z \right) \end{aligned}$$

Next you need to set the gradient equal to zero and solve the equations. This yields $y = 5, x = 3, z = -2$. Now to use the second derivative test, you assemble the Hessian matrix which is

$$\begin{pmatrix} \frac{5}{3} & -\frac{7}{3} & -\frac{4}{3} \\ -\frac{7}{3} & \frac{5}{3} & -\frac{4}{3} \\ -\frac{4}{3} & -\frac{4}{3} & \frac{2}{3} \end{pmatrix}.$$

Note that in this simple example, the Hessian matrix is constant and so all that is left is to consider the eigenvalues. Writing the characteristic equation and solving yields the eigenvalues are 2, -2, 4. Thus the given point is a saddle point.

Remember that all you care about is the sign of the eigenvalues. You don't have to find them exactly.

23.4 Exercises

1. Use the second derivative test on the critical points $(1, 1)$, and $(1, -1)$ for Example 23.3.5. The function is $6xy^2 - 2x^3 - 3x^4$.
2. Show the points $(\frac{1}{2}, -\frac{21}{4})$, $(0, -4)$, and $(1, -4)$ are critical points of the following function of two variables and classify them as local minima, local maxima or saddle points.

$$f(x, y) = -x^4 + 2x^3 + 39x^2 + 10yx^2 - 10yx - 40x - y^2 - 8y - 16.$$
3. Show the points $(\frac{1}{2}, -\frac{53}{12})$, $(0, -4)$, and $(1, -4)$ are critical points of the following function of two variables and classify them according to whether they are local minima, local maxima or saddle points.

$$f(x, y) = -3x^4 + 6x^3 + 37x^2 + 10yx^2 - 10yx - 40x - 3y^2 - 24y - 48.$$
4. Show the points $(\frac{1}{2}, \frac{37}{20})$, $(0, 2)$, and $(1, 2)$ are critical points of the following function of two variables and classify them according to whether they are local minima, local maxima or saddle points.

$$f(x, y) = 5x^4 - 10x^3 + 17x^2 - 6yx^2 + 6yx - 12x + 5y^2 - 20y + 20.$$
5. Show the points $(\frac{1}{2}, -\frac{17}{8})$, $(0, -2)$, and $(1, -2)$ are critical points of the following function of two variables and classify them according to whether they are local minima, local maxima or saddle points.

$$f(x, y) = 4x^4 - 8x^3 - 4yx^2 + 4yx + 8x - 4x^2 + 4y^2 + 16y + 16.$$
6. Find the critical points of the following function of three variables and classify them according to whether they are local minima, local maxima or saddle points.

$$f(x, y, z) = \frac{1}{3}x^2 + \frac{32}{3}x + \frac{4}{3} - \frac{16}{3}yx - \frac{58}{3}y - \frac{4}{3}zx - \frac{46}{3}z + \frac{1}{3}y^2 - \frac{4}{3}zy - \frac{5}{3}z^2.$$

7. Find the critical points of the following function of three variables and classify them according to whether they are local minima, local maxima or saddle points.

$$f(x, y, z) = -\frac{5}{3}x^2 + \frac{2}{3}x - \frac{2}{3} + \frac{8}{3}yx + \frac{2}{3}y + \frac{14}{3}zx - \frac{28}{3}z - \frac{5}{3}y^2 + \frac{14}{3}zy - \frac{8}{3}z^2.$$

8. Find the critical points of the following function of three variables and classify them according to whether they are local minima, local maxima or saddle points.

$$f(x, y, z) = -\frac{11}{3}x^2 + \frac{40}{3}x - \frac{56}{3} + \frac{8}{3}yx + \frac{10}{3}y - \frac{4}{3}zx + \frac{22}{3}z - \frac{11}{3}y^2 - \frac{4}{3}zy - \frac{5}{3}z^2.$$

9. Find the critical points of the following function of three variables and classify them according to whether they are local minima, local maxima or saddle points.

$$f(x, y, z) = -\frac{2}{3}x^2 + \frac{28}{3}x + \frac{37}{3} + \frac{14}{3}yx + \frac{10}{3}y - \frac{4}{3}zx - \frac{26}{3}z - \frac{2}{3}y^2 - \frac{4}{3}zy + \frac{7}{3}z^2.$$

10. *Show that if f has a critical point and some eigenvalue of the Hessian matrix is positive, then there exists a direction in which when f is evaluated on the line through the critical point having this direction, the resulting function of one variable has a local minimum. State and prove a similar result in the case where some eigenvalue of the Hessian matrix is negative.

11. Suppose $\mu = 0$ but there are negative eigenvalues of the Hessian at a critical point. Show by giving examples that the second derivative tests fails.

12. Show that the points $(\frac{1}{2}, -\frac{9}{2})$, $(0, -5)$, and $(1, -5)$ are critical points of the following function of two variables and classify them as local minima, local maxima or saddle points.

$$f(x, y) = 2x^4 - 4x^3 + 42x^2 + 8yx^2 - 8yx - 40x + 2y^2 + 20y + 50.$$

13. Show that the points $(1, -\frac{11}{2})$, $(0, -5)$, and $(2, -5)$ are critical points of the following function of two variables and classify them as local minima, local maxima or saddle points.

$$f(x, y) = 4x^4 - 16x^3 - 4x^2 - 4yx^2 + 8yx + 40x + 4y^2 + 40y + 100.$$

14. Show that the points $(\frac{3}{2}, \frac{27}{20})$, $(0, 0)$, and $(3, 0)$ are critical points of the following function of two variables and classify them as local minima, local maxima or saddle points.

$$f(x, y) = 5x^4 - 30x^3 + 45x^2 + 6yx^2 - 18yx + 5y^2.$$

15. Find the critical points of the following function of three variables and classify them as local minima, local maxima or saddle points.

$$f(x, y, z) = \frac{10}{3}x^2 - \frac{44}{3}x + \frac{64}{3} - \frac{10}{3}yx + \frac{16}{3}y + \frac{2}{3}zx - \frac{20}{3}z + \frac{10}{3}y^2 + \frac{2}{3}zy + \frac{4}{3}z^2.$$

16. Find the critical points of the following function of three variables and classify them as local minima, local maxima or saddle points.

$$f(x, y, z) = -\frac{7}{3}x^2 - \frac{146}{3}x + \frac{83}{3} + \frac{16}{3}yx + \frac{4}{3}y - \frac{14}{3}zx + \frac{94}{3}z - \frac{7}{3}y^2 - \frac{14}{3}zy + \frac{8}{3}z^2.$$

17. Find the critical points of the following function of three variables and classify them as local minima, local maxima or saddle points.

$$f(x, y, z) = \frac{2}{3}x^2 + 4x + 75 - \frac{14}{3}yx - 38y - \frac{8}{3}zx - 2z + \frac{2}{3}y^2 - \frac{8}{3}zy - \frac{1}{3}z^2.$$

18. Find the critical points of the following function of three variables and classify them as local minima, local maxima or saddle points.

$$f(x, y, z) = 4x^2 - 30x + 510 - 2yx + 60y - 2zx - 70z + 4y^2 - 2zy + 4z^2.$$

19. Show that the critical points of the following function are points of the form, $(x, y, z) = (t, 2t^2 - 10t, -t^2 + 5t)$ for $t \in \mathbb{R}$ and classify them as local minima, local maxima or saddle points.

$$f(x, y, z) = -\frac{1}{6}x^4 + \frac{5}{3}x^3 - \frac{25}{6}x^2 + \frac{10}{3}yx^2 - \frac{50}{3}yx + \frac{19}{3}zx^2 - \frac{95}{3}zx - \frac{5}{3}y^2 - \frac{10}{3}zy - \frac{1}{6}z^2.$$

20. Show that the critical points of the following function are

$$(0, -3, 0), (2, -3, 0), \text{ and } \left(1, -3, -\frac{1}{3}\right)$$

and classify them as local minima, local maxima or saddle points.

$$f(x, y, z) = -\frac{3}{2}x^4 + 6x^3 - 6x^2 + zx^2 - 2zx - 2y^2 - 12y - 18 - \frac{3}{2}z^2.$$

21. Show that the critical points of the function $f(x, y, z) = -2yx^2 - 6yx - 4zx^2 - 12zx + y^2 + 2yz$ are points of the form, $(x, y, z) = (t, 2t^2 + 6t, -t^2 - 3t)$ for $t \in \mathbb{R}$ and classify them as local minima, local maxima or saddle points.

22. Show that the critical points of the function

$$f(x, y, z) = \frac{1}{2}x^4 - 4x^3 + 8x^2 - 3zx^2 + 12zx + 2y^2 + 4y + 2 + \frac{1}{2}z^2.$$

are $(0, -1, 0)$, $(4, -1, 0)$, and $(2, -1, -12)$ and classify them as local minima, local maxima or saddle points.

23. Suppose $f(x, y)$, a function of two variables defined on all \mathbb{R}^p has all directional derivatives at $(0, 0)$ and they are all equal to 0 there. Suppose also that for $h(t) \equiv f(tu, tv)$ and (u, v) a unit vector, it follows that $h''(0) > 0$. By the one variable second derivative test, this implies that along every straight line through $(0, 0)$ the function restricted to this line has a local minimum at $(0, 0)$. Can it be concluded that f has a local minimum at $(0, 0)$. In other words, can you conclude a point is a local minimum if it appears to be so along every straight line through the point? **Hint:** Consider $f(x, y) = x^2 + y^2$ for (x, y) not on the curve $y = x^2$ for $x \neq 0$ and on this curve, let $f = -1$.

23.5 Lagrange Multipliers

Lagrange multipliers are used to solve extremum problems for a function defined on a level set of another function. This is the typical situation in optimization. You have a constraint on the variables and subject to this constraint, you are trying to maximize or minimize some function. It is the constraint which makes the problem interesting. For example, suppose you want to maximize xy given that $x + y = 4$. Solve for one of the variables say y , in the constraint equation $x + y = 4$ or $x + y - 4 = 0$ to find $y = 4 - x$. Then substitute this in to the function you are trying to maximize and take a derivative. The difficulty comes when you can't solve for one of the variables in the constraint or perhaps you could, but it would be inconvenient to do so.

In general, you want to maximize (minimize) $f(x, y, z)$ subject to the constraint

$$g(x, y, z) = 0.$$

Just because you can't algebraically solve for one of the variables, doesn't mean the relation does not define one of the variables in terms of the others. Say $z = z(x, y)$ near a point (x_0, y_0, z_0) on the constraint surface where the maximum or minimum exists. Then you could consider the unconstrained problem

$$(x, y) \rightarrow f(x, y, z(x, y))$$

and you would expect its partial derivatives to be 0 at the point of interest. By the chain rule (never mind the mathematical questions on existence), at this special point,

$$f_x + f_z z_x = 0, \quad f_y + f_z z_y = 0$$

By the process of implicit differentiation applied to $g(x, y, z) = 0$,

$$z_x = -\frac{g_x}{g_z}, \quad z_y = -\frac{g_y}{g_z}$$

Thus,

$$f_x = f_z \frac{g_x}{g_z} = \left(\frac{f_z}{g_z} \right) g_x, \quad f_y = f_z \frac{g_y}{g_z} = \left(\frac{f_z}{g_z} \right) g_y, \quad f_z = \left(\frac{f_z}{g_z} \right) g_z$$

So letting $\lambda = \frac{f_z(x_0, y_0, z_0)}{g_z(x_0, y_0, z_0)}$, it follows that at this point

$$\nabla f(x_0, y_0, z_0) = \lambda \nabla g(x_0, y_0, z_0)$$

The situation in which it is x or y that is a function of the other variables is exactly similar. Also, if there are more or fewer variables there is no difference in the argument. This λ is called a **Lagrange multiplier** after Lagrange who considered such problems in the 1700's.

Example 23.5.1 Maximize xyz subject to $x^2 + y^2 + z^2 = 27$.

Here $f(x, y, z) = xyz$ while $g(x, y, z) = x^2 + y^2 + z^2 - 27$. Then $\nabla g(x, y, z) = (2x, 2y, 2z)$ and $\nabla f(x, y, z) = (yz, xz, xy)$. Then at the point which maximizes this function¹,

$$(yz, xz, xy) = \lambda (2x, 2y, 2z).$$

Therefore, each of $2\lambda x^2, 2\lambda y^2, 2\lambda z^2$ equals xyz . It follows that at any point which maximizes xyz , $|x| = |y| = |z|$. Therefore, the only candidates for the point where the maximum occurs are

$$(3, 3, 3), (-3, -3, 3), (-3, 3, 3)$$

etc. The maximum occurs at $(3, 3, 3)$ which can be verified by plugging in to the function which is being maximized.

The method of Lagrange multipliers allows you to consider maximization of functions defined on closed and bounded sets. Recall that any continuous function defined on a closed and bounded set has a maximum and a minimum on the set. Candidates for the extremum on the interior of the set can be located by setting the gradient equal to zero. The consideration of the boundary can then sometimes be handled with the method of Lagrange multipliers.

¹There exists such a point because the sphere is closed and bounded.

Example 23.5.2 Maximize $f(x, y) = xy + y$ subject to the constraint, $x^2 + y^2 \leq 1$.

Here I know there is a maximum because the set is the closed disk, a closed and bounded set. Therefore, it is just a matter of finding it. Look for singular points on the interior of the circle. $\nabla f(x, y) = (y, x + 1) = (0, 0)$. There are no points on the interior of the circle where the gradient equals zero. Therefore, the maximum occurs on the boundary of the circle. That is, the problem reduces to maximizing $xy + y$ subject to $x^2 + y^2 = 1$. From the above,

$$(y, x + 1) - \lambda(2x, 2y) = 0.$$

Hence $y^2 - 2\lambda xy = 0$ and $x(x + 1) - 2\lambda xy = 0$ so $y^2 = x(x + 1)$. Therefore from the constraint, $x^2 + x(x + 1) = 1$ and the solution is $x = -1, x = \frac{1}{2}$. Then the candidates for a solution are $(-1, 0), \left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right), \left(\frac{1}{2}, -\frac{\sqrt{3}}{2}\right)$. Then

$$f(-1, 0) = 0, f\left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right) = \frac{3\sqrt{3}}{4}, f\left(\frac{1}{2}, -\frac{\sqrt{3}}{2}\right) = -\frac{3\sqrt{3}}{4}.$$

It follows the maximum value of this function is $\frac{3\sqrt{3}}{4}$ and it occurs at $\left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right)$. The minimum value is $-\frac{3\sqrt{3}}{4}$ and it occurs at $\left(\frac{1}{2}, -\frac{\sqrt{3}}{2}\right)$.

Example 23.5.3 Find candidates for the maximum and minimum values of the function $f(x, y) = xy - x^2$ on the set $\{(x, y) : x^2 + 2xy + y^2 \leq 4\}$.

First, the only point where ∇f equals zero is $(x, y) = (0, 0)$ and this is in the desired set. In fact it is an interior point of this set. This takes care of the interior points. What about those on the boundary $x^2 + 2xy + y^2 = 4$? The problem is to maximize $xy - x^2$ subject to the constraint, $x^2 + 2xy + y^2 = 4$. The Lagrangian is $xy - x^2 - \lambda(x^2 + 2xy + y^2 - 4)$ and this yields the following system.

$$\begin{aligned} y - 2x - \lambda(2x + 2y) &= 0 \\ x - 2\lambda(x + y) &= 0 \\ x^2 + 2xy + y^2 &= 4 \end{aligned}$$

From the first two equations,

$$(2 + 2\lambda)x - (1 - 2\lambda)y = 0, (1 - 2\lambda)x - 2\lambda y = 0$$

Since not both x and y equal zero, it follows

$$\det \begin{pmatrix} 2 + 2\lambda & 2\lambda - 1 \\ 1 - 2\lambda & -2\lambda \end{pmatrix} = 0$$

which yields $\lambda = 1/8$. Therefore, $y = 3x$. From the constraint equation $x^2 + 2x(3x) + (3x)^2 = 4$ and so $x = \frac{1}{2}$ or $-\frac{1}{2}$. Now since $y = 3x$, the points of interest on the boundary of this set are

$$\left(\frac{1}{2}, \frac{3}{2}\right), \text{ and } \left(-\frac{1}{2}, -\frac{3}{2}\right). \quad (23.1)$$

$$f\left(\frac{1}{2}, \frac{3}{2}\right) = \left(\frac{1}{2}\right)\left(\frac{3}{2}\right) - \left(\frac{1}{2}\right)^2 = \frac{1}{2}$$

$$f\left(-\frac{1}{2}, -\frac{3}{2}\right) = \left(-\frac{1}{2}\right)\left(-\frac{3}{2}\right) - \left(-\frac{1}{2}\right)^2 = \frac{1}{2}$$

Thus the candidates for maximum and minimum are $(\frac{1}{2}, \frac{3}{2})$, $(0, 0)$, and $(-\frac{1}{2}, -\frac{3}{2})$. Therefore it appears that $(0, 0)$ yields a minimum and either $(\frac{1}{2}, \frac{3}{2})$ or $(-\frac{1}{2}, -\frac{3}{2})$ yields a maximum. However, this is a little misleading. How do you even know a maximum or a minimum exists? The set $x^2 + 2xy + y^2 \leq 4$ is an unbounded set which lies between the two lines $x + y = 2$ and $x + y = -2$. In fact there is no minimum. For example, take $x = 100, y = -98$. Then $xy - x^2 = x(y - x) = 100(-98 - 100)$ which is a large negative number much less than 0, the answer for the point $(0, 0)$.

There are no magic bullets here. It was still required to solve a system of nonlinear equations to get the answer. However, it does often help to do it this way.

A nice observation in the case that the function f , which you are trying to maximize, and the function g , which defines the constraint, are functions of two or three variables is the following.

At points of interest,

$$\nabla f \times \nabla g = \mathbf{0}$$

This follows from the above because at these points,

$$\nabla f = \lambda \nabla g$$

so the angle between the two vectors ∇f and ∇g is either 0 or π . Therefore, the sine of this angle equals 0. By the geometric description of the cross product, this implies the cross product equals 0. Here is an example.

Example 23.5.4 Minimize $f(x, y) = xy - x^2$ on the set

$$\{(x, y) : x^2 + 2xy + y^2 = 4\}$$

Using the observation about the cross product, and letting $f(x, y, z) = f(x, y)$ with a similar convention for g , $\nabla f = (y - 2x, x, 0)$, $\nabla g = (2x + 2y, 2x + 2y, 0)$ and so

$$\begin{aligned} & (y - 2x, x, 0) \times (2x + 2y, 2x + 2y, 0) \\ &= (0, 0, (y - 2x)(2x + 2y) - x(2x + 2y)) = 0 \end{aligned}$$

Thus there are two equations, $x^2 + 2xy + y^2 = 4$ and $4xy - 2y^2 + 6x^2 = 0$. Solving these two yields the points of interest $(-\frac{1}{2}, -\frac{3}{2})$, $(\frac{1}{2}, \frac{3}{2})$. Both give the same value for f a maximum.

The above generalizes to a general procedure which is described in the following major Theorem. All correct proofs of this theorem will involve some appeal to the implicit function theorem or to fundamental existence theorems from differential equations. A complete proof is very fascinating but it will not come cheap. Good advanced calculus books will usually give a correct proof. If you are interested, there is a complete proof later. First here is a simple definition explaining one of the terms in the statement of this theorem.

Definition 23.5.5 Let A be an $m \times n$ matrix. A submatrix is any matrix which can be obtained from A by deleting some rows and some columns.

Theorem 23.5.6 Let U be an open subset of \mathbb{R}^n and let $f : U \rightarrow \mathbb{R}$ be a C^1 function. Then if $\mathbf{x}_0 \in U$, has the property that

$$g_i(\mathbf{x}_0) = 0, \quad i = 1, \dots, m, \quad g_i \text{ a } C^1 \text{ function}, \quad (23.2)$$

and \mathbf{x}_0 is either a local maximum or local minimum of f on the intersection of the level sets just described, and if some $m \times m$ submatrix of

$$Dg(\mathbf{x}_0) \equiv \begin{pmatrix} g_{1x_1}(\mathbf{x}_0) & g_{1x_2}(\mathbf{x}_0) & \cdots & g_{1x_n}(\mathbf{x}_0) \\ \vdots & \vdots & & \vdots \\ g_{mx_1}(\mathbf{x}_0) & g_{mx_2}(\mathbf{x}_0) & \cdots & g_{mx_n}(\mathbf{x}_0) \end{pmatrix}$$

has nonzero determinant, then there exist scalars, $\lambda_1, \dots, \lambda_m$ such that

$$\begin{pmatrix} f_{x_1}(\mathbf{x}_0) \\ \vdots \\ f_{x_n}(\mathbf{x}_0) \end{pmatrix} = \lambda_1 \begin{pmatrix} g_{1x_1}(\mathbf{x}_0) \\ \vdots \\ g_{1x_n}(\mathbf{x}_0) \end{pmatrix} + \cdots + \lambda_m \begin{pmatrix} g_{mx_1}(\mathbf{x}_0) \\ \vdots \\ g_{mx_n}(\mathbf{x}_0) \end{pmatrix} \quad (23.3)$$

holds.

To help remember how to use 23.3, do the following. First write the Lagrangian,

$$L = f(\mathbf{x}) - \sum_{i=1}^m \lambda_i g_i(\mathbf{x})$$

and then proceed to take derivatives with respect to each of the components of \mathbf{x} and also derivatives with respect to each λ_i and set all of these equations equal to 0. The formula 23.3 is what results from taking the derivatives of L with respect to the components of \mathbf{x} . When you take the derivatives with respect to the Lagrange multipliers, and set what results equal to 0, you just pick up the constraint equations. This yields $n + m$ equations for the $n + m$ unknowns $x_1, \dots, x_n, \lambda_1, \dots, \lambda_m$. Then you proceed to look for solutions to these equations. Of course these might be impossible to find using methods of algebra, but you just do your best and hope it will work out.

Example 23.5.7 Minimize xyz subject to the constraints $x^2 + y^2 + z^2 = 4$ and $x - 2y = 0$.

Form the Lagrangian,

$$L = xyz - \lambda (x^2 + y^2 + z^2 - 4) - \mu (x - 2y)$$

and proceed to take derivatives with respect to every possible variable, leading to the following system of equations.

$$\begin{aligned} yz - 2\lambda x - \mu &= 0 \\ xz - 2\lambda y + 2\mu &= 0 \\ xy - 2\lambda z &= 0 \\ x^2 + y^2 + z^2 &= 4 \\ x - 2y &= 0 \end{aligned}$$

Now you have to find the solutions to this system of equations. In general, this could be very hard or even impossible. If $\lambda = 0$, then from the third equation, either x or y must equal 0. Therefore, from the first two equations, $\mu = 0$ also. If $\mu = 0$ and $\lambda \neq 0$, then from the first two equations, $xyz = 2\lambda x^2$ and $xyz = 2\lambda y^2$ and so either $x = y$ or $x = -y$, which requires that both x and y equal zero thanks to the last equation. But then from the fourth

equation, $z = \pm 2$ and now this contradicts the third equation. Thus μ and λ are either both equal to zero or neither one is and the expression, xyz equals zero in this case. However, I know this is not the best value for a minimizer because I can take $x = 2\sqrt{\frac{3}{5}}, y = \sqrt{\frac{3}{5}}$, and $z = -1$. This satisfies the constraints and the product of these numbers equals a negative number. Therefore, both μ and λ must be non zero. Now use the last equation eliminate x and write the following system.

$$\begin{aligned} 5y^2 + z^2 &= 4 \\ y^2 - \lambda z &= 0 \\ yz - \lambda y + \mu &= 0 \\ yz - 4\lambda y - \mu &= 0 \end{aligned}$$

From the last equation, $\mu = (yz - 4\lambda y)$. Substitute this into the third and get

$$\begin{aligned} 5y^2 + z^2 &= 4 \\ y^2 - \lambda z &= 0 \\ yz - \lambda y + yz - 4\lambda y &= 0 \end{aligned}$$

$y = 0$ will not yield the minimum value from the above example. Therefore, divide the last equation by y and solve for λ to get $\lambda = (2/5)z$. Now put this in the second equation to conclude

$$\begin{aligned} 5y^2 + z^2 &= 4 \\ y^2 - (2/5)z^2 &= 0 \end{aligned}$$

a system which is easy to solve. Thus $y^2 = 8/15$ and $z^2 = 4/3$. Therefore, candidates for minima are $\left(2\sqrt{\frac{8}{15}}, \sqrt{\frac{8}{15}}, \pm\sqrt{\frac{4}{3}}\right)$, and $\left(-2\sqrt{\frac{8}{15}}, -\sqrt{\frac{8}{15}}, \pm\sqrt{\frac{4}{3}}\right)$, a choice of 4 points to check. Clearly the one which gives the smallest value is

$$\left(2\sqrt{\frac{8}{15}}, \sqrt{\frac{8}{15}}, -\sqrt{\frac{4}{3}}\right)$$

or $\left(-2\sqrt{\frac{8}{15}}, -\sqrt{\frac{8}{15}}, -\sqrt{\frac{4}{3}}\right)$ and the minimum value of the function subject to the constraints is $-\frac{2}{5}\sqrt{30} - \frac{2}{3}\sqrt{3}$.

You should rework this problem first solving the second easy constraint for x and then producing a simpler problem involving only the variables y and z .

23.6 Exercises

1. Maximize $x + y + z$ subject to the constraint $x^2 + y^2 + z^2 = 3$.
2. Minimize $2x - y + z$ subject to the constraint $2x^2 + y^2 + z^2 = 36$.
3. Minimize $x + 3y - z$ subject to the constraint $2x^2 + y^2 - 2z^2 = 36$ if possible. Note there is no guaranty this function has either a maximum or a minimum. Determine whether there exists a minimum also.
4. Find the dimensions of the largest rectangle which can be inscribed in a circle of radius r .

5. Maximize $2x + y$ subject to the condition that $\frac{x^2}{4} + \frac{y^2}{9} \leq 1$.
6. Maximize $x + 2y$ subject to the condition that $x^2 + \frac{y^2}{9} \leq 1$.
7. Maximize $x + y$ subject to the condition that $x^2 + \frac{y^2}{9} + z^2 \leq 1$.
8. Minimize $x + y + z$ subject to the condition that $x^2 + \frac{y^2}{9} + z^2 \leq 1$.
9. Find the points on $y^2x = 16$ which are closest to $(0, 0)$.
10. Find the points on $\sqrt{2}y^2x = 1$ which are closest to $(0, 0)$.
11. Find points on $xy = 4$ farthest from $(0, 0)$ if any exist. If none exist, tell why. What does this say about the method of Lagrange multipliers?
12. A can is supposed to have a volume of 36π cubic centimeters. Find the dimensions of the can which minimizes the surface area.
13. A can is supposed to have a volume of 36π cubic centimeters. The top and bottom of the can are made of tin costing 4 cents per square centimeter and the sides of the can are made of aluminum costing 5 cents per square centimeter. Find the dimensions of the can which minimizes the cost.
14. Minimize and maximize $\sum_{j=1}^n x_j$ 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.
15. Find the point (x, y, z) on the level surface $4x^2 + y^2 - z^2 = 1$ which is closest to $(0, 0, 0)$.
16. A curve is formed from the intersection of the plane, $2x + y + z = 3$ 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. Find the point on the plane, $2x + 3y + z = 4$ which is closest to the point $(1, 2, 3)$.
19. Let $A = (A_{ij})$ be an $n \times n$ matrix which is symmetric. Thus $A_{ij} = A_{ji}$ and recall $(A\mathbf{x})_i = A_{ij}x_j$ where as usual, sum over the repeated index. Show that $\frac{\partial}{\partial x_k} (A_{ij}x_jx_i) = 2A_{ik}x_k$. 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 $A\mathbf{x} = \lambda\mathbf{x}$. Thus λ is an eigenvalue of the matrix A .
20. Here are two lines.

$$\mathbf{x} = (1 + 2t, 2 + t, 3 + t)^T$$
and $\mathbf{x} = (2 + s, 1 + 2s, 1 + 3s)^T$. Find points \mathbf{p}_1 on the first line and \mathbf{p}_2 on the second with the property that $|\mathbf{p}_1 - \mathbf{p}_2|$ is at least as small as the distance between any other pair of points, one chosen on one line and the other on the other line.
21. * Find points on the circle of radius r for the largest triangle which can be inscribed in it.

22. Find the point on the intersection of $z = x^2 + y^2$ and $x + y + z = 1$ which is closest to $(0, 0, 0)$.
23. Minimize xyz subject to the constraints $x^2 + y^2 + z^2 = r^2$ and $x - y = 0$.
24. Let n be a positive integer. Find n numbers whose sum is $8n$ and the sum of the squares is as small as possible.
25. Find the point on the level surface $2x^2 + xy + z^2 = 16$ which is closest to $(0, 0, 0)$.
26. Find the point on $x^2 + y^2 + z^2 = 1$ closest to the plane $x + y + z = 10$.
27. Find the point on $\frac{x^2}{4} + \frac{y^2}{9} + z^2 = 1$ closest to the plane $x + y + z = 10$.
28. Let x_1, \dots, x_5 be 5 positive numbers. Maximize their product subject to the constraint that

$$x_1 + 2x_2 + 3x_3 + 4x_4 + 5x_5 = 300.$$

29. Let $f(x_1, \dots, x_n) = x_1^n x_2^{n-1} \cdots x_n^1$. Then f achieves a maximum on the set $S \equiv$

$$\left\{ \mathbf{x} \in \mathbb{R}^n : \sum_{i=1}^n ix_i = 1, \text{ each } x_i \geq 0 \right\}$$

If $\mathbf{x} \in S$ is the point where this maximum is achieved, find x_1/x_n .

30. * Let (x, y) be a point on the ellipse, $x^2/a^2 + y^2/b^2 = 1$ which is in the first quadrant. Extend the tangent line through (x, y) till it intersects the x and y axes and let $A(x, y)$ denote the area of the triangle formed by this line and the two coordinate axes. Find the minimum value of the area of this triangle as a function of a and b .

31. Maximize $\prod_{i=1}^n x_i^2$

$$(\equiv x_1^2 \times x_2^2 \times x_3^2 \times \cdots \times x_n^2)$$

subject to the constraint, $\sum_{i=1}^n x_i^2 = r^2$. Show that the maximum is $(r^2/n)^n$. 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.

32. Maximize $x^2 y^2$ subject to the constraint

$$\frac{x^{2p}}{p} + \frac{y^{2q}}{q} = r^2$$

where p, q are real numbers larger than 1 which have the property that

$$\frac{1}{p} + \frac{1}{q} = 1$$

show that the maximum is achieved when $x^{2p} = y^{2q}$ and equals r^2 . Now conclude that if $x, y > 0$, then

$$xy \leq \frac{x^p}{p} + \frac{y^q}{q}$$

and there are values of x and y where this inequality is an equation.

33. The area of the ellipse $x^2/a^2 + y^2/b^2 \leq 1$ is πab which is given to equal π . The length of the ellipse is $\int_0^{2\pi} \sqrt{a^2 \sin^2(t) + b^2 \cos^2(t)} dt$. Find a, b such that the ellipse having this volume is as short as possible.
34. Consider the closed region in the xy plane which lies between the curve $y = \sqrt{1-x^2}$ and $y = 0$. Find the maximum and minimum values of the function $x^2 + x + y^2 - y$ on this region. **Hint:** First observe that there is a solution because the region is compact. Next look for candidates for the extreme point on the interior. When this is done, look for candidates on the boundary. Note that the boundary of the region does not come as the level surface of a C^1 function. The method does not apply to the corners of this region, the points $(1, 0)$ and $(-1, 0)$. Therefore, you need to consider these points also.
35. To see why the method works with more than one constraint, suppose you have the problem to maximize $f(x, y, z)$ with the constraints

$$g_1(x, y, z) = 0, g_2(x, y, z) = 0$$

Then the two constraints likely define a curve of intersection. Say $z = z(x), y = y(x)$. At the point where a maximum or minimum occurs, explain why

$$\begin{aligned} f_x + f_y y_x + f_z z_x &= 0 \\ g_{1x} + g_{1y} y_x + g_{1z} z_x &= 0 \\ g_{2x} + g_{2y} y_x + g_{2z} z_x &= 0 \end{aligned}$$

This is a system of equations having nonzero solution $(1, y_x, z_x)$. Thus the matrix of coefficients has no inverse. Thus the rows are dependent. If $\nabla g_1, \nabla g_2$ are independent, $\nabla f = \lambda_1 \nabla g_1 + \lambda_2 \nabla g_2$ for some scalars λ_i . Other situations are similar but to do this in full generality, see the appendix on implicit function theorem.

36. Suppose you wish to maximize(minimize) $f(\mathbf{x})$ subject to $g(\mathbf{x}) = 0$ where $\mathbf{x} \in \mathbb{R}^n, n \geq 1$. Say $\mathbf{x} = (x_1, \dots, x_{n-1}, x_n)$ and at a point \mathbf{x}_0 where the minimum of maximum occurs, you have $\mathbf{x}_n = x_n(x_1, \dots, x_{n-1}), g_{x_n} \neq 0$. The situation is the same if $g(\mathbf{x}) = 0$ defines one of the other variables as a function of the remaining variables. Then, assuming all functions are C^1 , (See appendix on implicit function theorem for this.) explain why you have for each $x_i, i \leq n-1$ at the point \mathbf{x}_0

$$\begin{aligned} f_{x_i} + f_{x_n} \frac{\partial x_n}{\partial x_i} &= 0 \\ g_{x_i} + g_{x_n} \frac{\partial x_n}{\partial x_i} &= 0 \end{aligned}$$

Thus

$$f_{x_i} = -f_{x_n} \left(\frac{-g_{x_i}}{g_{x_n}} \right) = \left(\frac{f_{x_n}}{g_{x_n}} \right) g_{x_i}$$

Explain why the gradient of f equals a multiple of the gradient of g at the point where the local extreme value occurs.

23.7 Proof of the Second Derivative Test*

A version of the following theorem is due to Lagrange, about 1790. The proof is given earlier. See 5.15.1 on Page 152. It is stated here for convenience.

Theorem 23.7.1 *Suppose f has $n+1$ derivatives on an interval (a, b) and let $c \in (a, b)$. Then if $x \in (a, b)$, there exists ξ between c and x such that*

$$f(x) = f(c) + \sum_{k=1}^n \frac{f^{(k)}(c)}{k!} (x-c)^k + \frac{f^{(n+1)}(\xi)}{(n+1)!} (x-c)^{n+1}.$$

(In this formula, the symbol $\sum_{k=1}^0 a_k$ will denote the number 0.)

Definition 23.7.2 *The matrix $\left(\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}) \right)$ is called the Hessian matrix, denoted by $H(\mathbf{x})$.*

Now recall the Taylor formula with the Lagrange form of the remainder.

Theorem 23.7.3 *Let $h : (-\delta, 1 + \delta) \rightarrow \mathbb{R}$ have $m+1$ derivatives. Then there exists $t \in (0, 1)$ such that*

$$h(1) = h(0) + \sum_{k=1}^m \frac{h^{(k)}(0)}{k!} + \frac{h^{(m+1)}(t)}{(m+1)!}.$$

Now let $f : U \rightarrow \mathbb{R}$ where U is an open subset of \mathbb{R}^p . Suppose $f \in C^2(U)$. Let $\mathbf{x} \in U$ and let $r > 0$ be such that

$$B(\mathbf{x}, r) \subseteq U.$$

Then for $\|\mathbf{v}\| < r$ consider

$$f(\mathbf{x} + t\mathbf{v}) - f(\mathbf{x}) \equiv h(t)$$

for $t \in [0, 1]$. Then from Taylor's theorem for the case where $m = 2$ and the chain rule,

$$h'(t) = \sum_i \frac{\partial f}{\partial x_i}(\mathbf{x} + t\mathbf{v}) v_i, h''(t) = \sum_j \sum_i \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x} + t\mathbf{v}) v_i v_j.$$

Thus

$$h''(t) = \mathbf{v}^T H(\mathbf{x} + t\mathbf{v}) \mathbf{v}.$$

From Theorem 23.7.3 there exists $t \in (0, 1)$ such that

$$f(\mathbf{x} + \mathbf{v}) = f(\mathbf{x}) + \frac{\partial f}{\partial x_i}(\mathbf{x}) v_i + \frac{1}{2} \mathbf{v}^T H(\mathbf{x} + t\mathbf{v}) \mathbf{v}$$

By the continuity of the second partial derivative

$$f(\mathbf{x} + \mathbf{v}) = f(\mathbf{x}) + \nabla f(\mathbf{x}) \cdot \mathbf{v} + \frac{1}{2} \mathbf{v}^T H(\mathbf{x}) \mathbf{v} + \frac{1}{2} (\mathbf{v}^T (H(\mathbf{x} + t\mathbf{v}) - H(\mathbf{x})) \mathbf{v}) \quad (23.4)$$

where the last term satisfies

$$\lim_{|\mathbf{v}| \rightarrow 0} \frac{1}{2} \frac{(\mathbf{v}^T (H(\mathbf{x} + t\mathbf{v}) - H(\mathbf{x})) \mathbf{v})}{|\mathbf{v}|^2} = 0 \quad (23.5)$$

because of the continuity of the entries of $H(\mathbf{x})$.

Theorem 23.7.4 Suppose \mathbf{x} is a critical point for f . That is, suppose $\frac{\partial f}{\partial x_i}(\mathbf{x}) = 0$ for each i . Then if $H(\mathbf{x})$ has all positive eigenvalues, \mathbf{x} is a local minimum. If $H(\mathbf{x})$ has all negative eigenvalues, then \mathbf{x} is a local maximum. If $H(\mathbf{x})$ has a positive eigenvalue, then there exists a direction in which f has a local minimum at \mathbf{x} , while if $H(\mathbf{x})$ has a negative eigenvalue, there exists a direction in which f has a local maximum at \mathbf{x} .

Proof: Since $\nabla f(\mathbf{x}) = \mathbf{0}$, formula 23.4 implies

$$f(\mathbf{x} + \mathbf{v}) = f(\mathbf{x}) + \frac{1}{2} \mathbf{v}^T H(\mathbf{x}) \mathbf{v} + \frac{1}{2} (\mathbf{v}^T (H(\mathbf{x} + t\mathbf{v}) - H(\mathbf{x})) \mathbf{v}) \quad (23.6)$$

and by continuity of the second derivatives, these mixed second derivatives are equal and so $H(\mathbf{x})$ is a symmetric matrix. Thus, by Theorem 19.8.6, $H(\mathbf{x})$ has all real eigenvalues and can be diagonalized with an orthogonal matrix U . Suppose first that $H(\mathbf{x})$ has all positive eigenvalues and that all are larger than $\delta^2 > 0$.

$$\mathbf{u}^T H(\mathbf{x}) \mathbf{u} = \mathbf{u}^T U D U^T \mathbf{u} = (U\mathbf{u})^T D (U\mathbf{u}) \geq \delta^2 |U\mathbf{u}|^2 = \delta^2 |\mathbf{u}|^2$$

By continuity of H , if \mathbf{v} is small enough,

$$f(\mathbf{x} + \mathbf{v}) \geq f(\mathbf{x}) + \frac{1}{2} \delta^2 |\mathbf{v}|^2 - \frac{1}{4} \delta^2 |\mathbf{v}|^2 = f(\mathbf{x}) + \frac{\delta^2}{4} |\mathbf{v}|^2.$$

This shows the first claim of the theorem. The second claim follows from similar reasoning or applying the above to $-f$.

Suppose $H(\mathbf{x})$ has a positive eigenvalue λ^2 . Then let \mathbf{v} be an eigenvector for this eigenvalue. Then from (23.6), replacing \mathbf{v} with $s\mathbf{v}$ and letting t depend on s ,

$$f(\mathbf{x} + s\mathbf{v}) = f(\mathbf{x}) + \frac{1}{2} s^2 \mathbf{v}^T H(\mathbf{x}) \mathbf{v} + \frac{1}{2} s^2 (\mathbf{v}^T (H(\mathbf{x} + ts\mathbf{v}) - H(\mathbf{x})) \mathbf{v})$$

which implies

$$\begin{aligned} f(\mathbf{x} + s\mathbf{v}) &= f(\mathbf{x}) + \frac{1}{2} s^2 \lambda^2 |\mathbf{v}|^2 + \frac{1}{2} s^2 (\mathbf{v}^T (H(\mathbf{x} + ts\mathbf{v}) - H(\mathbf{x})) \mathbf{v}) \\ &\geq f(\mathbf{x}) + \frac{1}{4} s^2 \lambda^2 |\mathbf{v}|^2 \end{aligned}$$

whenever s is small enough. Thus in the direction \mathbf{v} the function has a local minimum at \mathbf{x} . The assertion about the local maximum in some direction follows similarly. ■

Chapter 24

Implicit Function Theorem*

This chapter can be skipped and returned to later if desired. It contains the theoretical background for the method of Lagrange multipliers and other items.

First is the version involving the case where f is a function of two scalar variables and has values in \mathbb{R} . The partial derivatives D_1f, D_2f are just derivatives obtained from fixing y in the first case and x in the second. Thus $D_1f(x, y) \equiv \lim_{h \rightarrow 0} \frac{f(x+h, y) - f(x, y)}{h}$, similar for $D_2f(x, y)$. Continuity means that if $x_n \rightarrow x$, and $y_n \rightarrow y$, then $f(x_n, y_n) \rightarrow f(x, y)$.

Theorem 24.0.1 *Let $(x_0, y_0) \in I \times J$ where I, J are open intervals and suppose $f(x_0, y_0) = 0$ where D_1f, D_2f, f are continuous on $I \times J$. Also suppose $D_1f(x_0, y_0) \neq 0$. Then there exist open intervals $\hat{I} \subseteq I, \hat{J} \subseteq J$ with $(x_0, y_0) \in \hat{I} \times \hat{J}$ and a unique function $x: \hat{J} \rightarrow \hat{I}$ which has continuous derivative such that $f(x(y), y) = 0$ for all $y \in \hat{J}$.*

Proof: There exist \tilde{I}, \tilde{J} closed intervals contained in I, J respectively such that x_0 is in the interior of \tilde{I} , y_0 is in the interior of \tilde{J} and $D_1f(x, y), D_2f(x, y)$ are both nonzero on $\tilde{I} \times \tilde{J}$. This is by continuity of the partial derivatives. Then fixing $y \in \tilde{J}$, if $f(x_1, y) = f(x_2, y) = 0$, you would have, by the mean value theorem, $D_1f(z, y)(x_1 - x_2) = 0$ for some z between x_1 and x_2 . Hence $D_1f(z, y) = 0$ which does not happen. Hence there is at most one x for which $f(x, y) = 0$ for each $y \in \tilde{J}$.

claim: There exists open $\hat{J} \subseteq \tilde{J}$ such that \hat{J} contains y_0 and for $y \in \hat{J}$, the minimum of $x \rightarrow f^2(x, y)$ on $\tilde{I} \times \hat{J}$ occurs at an interior point of \tilde{I} .

Proof of claim: If not, there exists $y_n \rightarrow y_0$ and x_n an endpoint of \tilde{I} such that the minimum of $x \rightarrow f^2(x, y)$ occurs at x_n . One of these endpoints occurs infinitely often and so there is a subsequence, still called x_n which converges to an endpoint w of \tilde{I} . Then $f^2(x_0, y_n) \geq f^2(x_n, y_n)$ and so, by continuity, $f^2(x_0, y_0) \geq f^2(w, y_0)$ and so, there are two different points of \tilde{I} namely x_0, w with $f(x_0, y_0) = f(w, y_0) = 0$. This was just shown impossible. This proves the claim.

Let \hat{J} be as just described and let $x(y)$ be the point of \tilde{I} for which $f(x(y), y)$ minimizes $x \rightarrow f^2(x, y)$. Then fixing y , $2f(x(y), y)D_1f(x(y), y) = 0$ and so, $f(x(y), y) = 0$. It remains to verify that $y \rightarrow x(y)$ is differentiable. If $y \in \hat{J}$, then if $|h|$ is small enough, both $y, y+h$ are in \hat{J} . Considering such small h ,

$$\begin{aligned} 0 &= f(x(y+h), y+h) - f(x(y), y) \\ &= f(x(y+h), y+h) - f(x(y), y+h) + f(x(y), y+h) - f(x(y), y) \\ &= D_1f(z(h), y+h)(x(y+h) - x(y)) + D_2f(x(y), y+h)h \end{aligned}$$

where w_h is between 0 and h and $z(h)$ is between $x(y)$ and $x(y+h)$. This is by the mean value theorem. Thus, the term on the right converges to 0 as $h \rightarrow 0$ and so it is also true that $x(y+h) - x(y) \rightarrow 0$ so $y \rightarrow x(y)$ is continuous. Then $z(h) \rightarrow x(y)$ also and so, by continuity you can take the limit as $h \rightarrow 0$ in

$$\frac{x(y+h) - x(y)}{h} = -\frac{D_2 f(x(y), y + w_h)}{D_1 f(z(h), y + h)}$$

and obtain $x'(y)$ exists and equals $-\frac{D_2 f(x(y), y)}{D_1 f(x(y), y)}$ which shows also that $y \rightarrow x'(y)$ is continuous by continuity of D_2 and D_1 . Now let $\hat{I} \equiv x(\hat{J})$. This smaller interval contains x_0 since $x(y_0) = x_0$ and is an open interval because $y \rightarrow x(y)$ is one to one and continuous. ■

The following is the implicit function theorem for functions of many variables. It is one of the greatest theorems in mathematics. The proof given here is like one found in one of Caratheodory's books on the calculus of variations and is a generalization of the above simpler case. I think this theorem was known to Weierstrass because it is used in a book by Bolza which is based on his lectures. The proof follows the easier one above and is not as elegant as some of the others which are based on a contraction mapping principle but it may be more accessible. However, it is an advanced topic. Don't waste your time with it unless you have first read and understood the earlier material on linear algebra. You will also need the extreme value theorem for a function of n variables and the chain rule of multi-variable calculus. First is an interesting proposition.

Proposition 24.0.2 *Suppose*

$$g : \overline{B(x_0, \delta)} \times \overline{B(y_0, \eta_0)} \rightarrow [0, \infty)$$

is continuous and $g(x_0, y_0) = 0$ and if $x \neq x_0, g(x, y_0) > 0$. Then there exists $\eta < \eta_0$ such that if $y \in B(y_0, \eta)$, then the function $x \rightarrow g(x, y)$ achieves its minimum on the open set $B(x_0, \delta)$.

Proof: If not, then there is a sequence $y_k \rightarrow y_0$ but the minimum of $x \rightarrow g(x, y_k)$ for $x \in \overline{B(x_0, \delta)}$ happens on $\partial B(x_0, \delta) \equiv \partial B \equiv \{x : |x - x_0| = \delta\}$ at x_k . Now ∂B is closed and bounded and so compact. Hence there is a subsequence, still denoted with subscript k such that $x_k \rightarrow x \in \partial B$ and $y_k \rightarrow y_0$.

By uniform continuity on the compact set and assumption, if k is large enough, for all $\hat{x} \in \overline{B(x_0, \delta)}$

$$\begin{aligned} & \max \left\{ |g(\hat{x}, y_0) - g(\hat{x}, y_k)| : \hat{x} \in \overline{B(x_0, \delta)} \right\} \\ & \leq \frac{1}{2} \min \{ |g(\hat{x}, y_0)| : \hat{x} \in \partial B \} \leq \frac{1}{2} g(x, y_0) \end{aligned}$$

Then

$$g(x_0, y_k) \geq g(x_k, y_k) > g(x_k, y_0) - \frac{1}{2} g(x, y_0)$$

Letting $k \rightarrow \infty, 0 \geq \frac{1}{2} g(x, y_0)$ contrary to $\hat{x} \rightarrow g(\hat{x}, y_0)$ is only 0 at x_0 . ■

Definition 24.0.3 *Suppose U is an open set in $\mathbb{R}^n \times \mathbb{R}^m$ and (x, y) will denote a typical point of $\mathbb{R}^n \times \mathbb{R}^m$ with $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$. Let $f : U \rightarrow \mathbb{R}^p$ be in $C^1(U)$. Then*

define

$$\begin{aligned} D_1 f(x, y) &\equiv \begin{pmatrix} f_{1,x_1}(x, y) & \cdots & f_{1,x_n}(x, y) \\ \vdots & & \vdots \\ f_{p,x_1}(x, y) & \cdots & f_{p,x_n}(x, y) \end{pmatrix}, \\ D_2 f(x, y) &\equiv \begin{pmatrix} f_{1,y_1}(x, y) & \cdots & f_{1,y_m}(x, y) \\ \vdots & & \vdots \\ f_{p,y_1}(x, y) & \cdots & f_{p,y_m}(x, y) \end{pmatrix}. \end{aligned}$$

Theorem 24.0.4 (implicit function theorem) Suppose U is an open set in $\mathbb{R}^n \times \mathbb{R}^m$. Let $f : U \rightarrow \mathbb{R}^n$ be in $C^1(U)$ and suppose

$$f(x_0, y_0) = 0, D_1 f(x_0, y_0)^{-1} \text{ exists.} \quad (24.1)$$

Then there exist positive constants, δ, η , such that for every $y \in B(y_0, \eta)$ there exists a unique $x(y) \in B(x_0, \delta)$ such that

$$f(x(y), y) = 0. \quad (24.2)$$

Furthermore, the mapping, $y \rightarrow x(y)$ is in $C^1(B(y_0, \eta))$.

Proof: Let

$$f(x, y) = (f_1(x, y) \ f_2(x, y) \ \cdots \ f_n(x, y))^T.$$

Define for $(x^1, \dots, x^n) \in \overline{B(x_0, \delta)}^n$ and $y \in B(y_0, \eta)$ the following matrix.

$$J(x^1, \dots, x^n, y) \equiv \begin{pmatrix} f_{1,x_1}(x^1, y) & \cdots & f_{1,x_n}(x^1, y) \\ \vdots & & \vdots \\ f_{n,x_1}(x^n, y) & \cdots & f_{n,x_n}(x^n, y) \end{pmatrix}. \quad (*)$$

Then by the assumption of continuity of all the partial derivatives, there exists $r > 0$ and $\delta_0, \eta_0 > 0$ such that if $\delta \leq \delta_0$ and $\eta \leq \eta_0$, it follows that for all $(x^1, \dots, x^n) \in \overline{B(x_0, \delta)}^n \equiv \overline{B(x_0, \delta)} \times \overline{B(x_0, \delta)} \times \cdots \times \overline{B(x_0, \delta)}$, and $y \in \overline{B(y_0, \eta)}$,

$$|\det(J(x^1, \dots, x^n, y))| > r > 0. \quad (24.3)$$

and $\overline{B(x_0, \delta_0)} \times \overline{B(y_0, \eta_0)} \subseteq U$. By continuity of all the partial derivatives and the extreme value theorem, it can also be assumed there exists a constant, K such that for all $(x, y) \in \overline{B(x_0, \delta_0)} \times \overline{B(y_0, \eta_0)}$ and $i = 1, 2, \dots, n$, the i^{th} row of $D_2 f(x, y)$, given by $D_2 f_i(x, y)$ satisfies

$$|D_2 f_i(x, y)| < K, \quad (24.4)$$

and for all $(x^1, \dots, x^n) \in \overline{B(x_0, \delta_0)}^n$ and $y \in \overline{B(y_0, \eta_0)}$ the i^{th} row of the matrix,

$$J(x^1, \dots, x^n, y)^{-1}$$

which equals $e_i^T (J(x^1, \dots, x^n, y)^{-1})$ satisfies

$$|e_i^T (J(x^1, \dots, x^n, y)^{-1})| < K. \quad (24.5)$$

(Recall that e_i is the column vector consisting of all zeros except for a 1 in the i^{th} position.)

To begin with it is shown that for a given $\mathbf{y} \in B(\mathbf{y}_0, \eta)$, $\eta \leq \eta_0$, there is at most one $\mathbf{x} \in B(\mathbf{x}_0, \delta_0)$ such that $\mathbf{f}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$.

Pick $\mathbf{y} \in B(\mathbf{y}_0, \eta)$ and suppose there exist $\mathbf{x}, \mathbf{z} \in \overline{B(\mathbf{x}_0, \delta)}$ such that

$$\mathbf{f}(\mathbf{x}, \mathbf{y}) = \mathbf{f}(\mathbf{z}, \mathbf{y}) = \mathbf{0}$$

Consider f_i and let

$$h(t) \equiv f_i(\mathbf{x} + t(\mathbf{z} - \mathbf{x}), \mathbf{y}).$$

Then $h(1) = h(0)$ and so by the mean value theorem, $h'(t_i) = 0$ for some $t_i \in (0, 1)$. Therefore, from the chain rule and for this value of t_i ,

$$h'(t_i) = \sum_{j=1}^n \frac{\partial}{\partial x_j} f_i(\mathbf{x} + t_i(\mathbf{z} - \mathbf{x}), \mathbf{y}) (z_j - x_j) = 0. \quad (24.6)$$

Then denote by \mathbf{x}^i the vector, $\mathbf{x} + t_i(\mathbf{z} - \mathbf{x})$. It follows from 24.6 that

$$J(\mathbf{x}^1, \dots, \mathbf{x}^n, \mathbf{y})(\mathbf{z} - \mathbf{x}) = \mathbf{0}$$

and so from 24.3 $\mathbf{z} - \mathbf{x} = \mathbf{0}$. (The matrix, in the above is invertible since its determinant is nonzero.) Now it will be shown that if η is chosen sufficiently small, then for all $\mathbf{y} \in B(\mathbf{y}_0, \eta)$, there exists a unique $\mathbf{x}(\mathbf{y}) \in B(\mathbf{x}_0, \delta)$ such that $\mathbf{f}(\mathbf{x}(\mathbf{y}), \mathbf{y}) = \mathbf{0}$.

Claim: If η is small enough, then the function, $\mathbf{x} \rightarrow h_{\mathbf{y}}(\mathbf{x}) \equiv |\mathbf{f}(\mathbf{x}, \mathbf{y})|^2$ achieves its minimum value on $\overline{B(\mathbf{x}_0, \delta)}$ at a point of $B(\mathbf{x}_0, \delta)$. This is Proposition 24.0.2.

Choose $\eta < \eta_0$ and also small enough that the above claim holds and let $\mathbf{x}(\mathbf{y})$ denote a point of $B(\mathbf{x}_0, \delta)$ at which the minimum of $h_{\mathbf{y}}$ on $\overline{B(\mathbf{x}_0, \delta)}$ is achieved. Since $\mathbf{x}(\mathbf{y})$ is an interior point, you can consider $h_{\mathbf{y}}(\mathbf{x}(\mathbf{y}) + t\mathbf{v})$ for $|t|$ small and conclude this function of t has a zero derivative at $t = 0$. Now

$$h_{\mathbf{y}}(\mathbf{x}(\mathbf{y}) + t\mathbf{v}) = \sum_{i=1}^n f_i^2(\mathbf{x}(\mathbf{y}) + t\mathbf{v}, \mathbf{y})$$

and so from the chain rule,

$$\frac{d}{dt} h_{\mathbf{y}}(\mathbf{x}(\mathbf{y}) + t\mathbf{v}) = \sum_{i=1}^n \sum_{j=1}^n 2f_i(\mathbf{x}(\mathbf{y}) + t\mathbf{v}, \mathbf{y}) \frac{\partial f_i(\mathbf{x}(\mathbf{y}) + t\mathbf{v}, \mathbf{y})}{\partial x_j} v_j.$$

Therefore, letting $t = 0$, it is required that for every \mathbf{v} ,

$$\sum_{i=1}^n \sum_{j=1}^n 2f_i(\mathbf{x}(\mathbf{y}), \mathbf{y}) \frac{\partial f_i(\mathbf{x}(\mathbf{y}), \mathbf{y})}{\partial x_j} v_j = 0.$$

In terms of matrices this reduces to $0 = 2\mathbf{f}(\mathbf{x}(\mathbf{y}), \mathbf{y})^T D_1 \mathbf{f}(\mathbf{x}(\mathbf{y}), \mathbf{y}) \mathbf{v}$ for every vector \mathbf{v} . Therefore, $\mathbf{0} = \mathbf{f}(\mathbf{x}(\mathbf{y}), \mathbf{y})^T D_1 \mathbf{f}(\mathbf{x}(\mathbf{y}), \mathbf{y})$. From 24.3, it follows $\mathbf{f}(\mathbf{x}(\mathbf{y}), \mathbf{y}) = \mathbf{0}$. This proves the existence of the function $\mathbf{y} \rightarrow \mathbf{x}(\mathbf{y})$ such that $\mathbf{f}(\mathbf{x}(\mathbf{y}), \mathbf{y}) = \mathbf{0}$ for all $\mathbf{y} \in B(\mathbf{y}_0, \eta)$.

It remains to verify this function is a C^1 function. To do this, let \mathbf{y}_1 and \mathbf{y}_2 be points of $B(\mathbf{y}_0, \eta)$. Then as before, consider the i^{th} component of \mathbf{f} and consider the same argument using the mean value theorem to write

$$\begin{aligned} 0 &= f_i(\mathbf{x}(\mathbf{y}_1), \mathbf{y}_1) - f_i(\mathbf{x}(\mathbf{y}_2), \mathbf{y}_2) \\ &= f_i(\mathbf{x}(\mathbf{y}_1), \mathbf{y}_1) - f_i(\mathbf{x}(\mathbf{y}_2), \mathbf{y}_1) + f_i(\mathbf{x}(\mathbf{y}_2), \mathbf{y}_1) - f_i(\mathbf{x}(\mathbf{y}_2), \mathbf{y}_2) \\ &= D_1 f_i(\mathbf{x}^i, \mathbf{y}_1)(\mathbf{x}(\mathbf{y}_1) - \mathbf{x}(\mathbf{y}_2)) + D_2 f_i(\mathbf{x}(\mathbf{y}_2), \mathbf{y}^i)(\mathbf{y}_1 - \mathbf{y}_2). \end{aligned} \quad (24.7)$$

where \mathbf{y}^i is a point on the line segment joining \mathbf{y}_1 and \mathbf{y}_2 . Thus from 24.4 and the Cauchy-Schwarz inequality,

$$|D_2 f_i(\mathbf{x}(\mathbf{y}_2), \mathbf{y}^i)(\mathbf{y}_1 - \mathbf{y}_2)| \leq K |\mathbf{y}_1 - \mathbf{y}_2|.$$

Therefore, letting $M(\mathbf{y}^1, \dots, \mathbf{y}^n) \equiv M$ denote the matrix having the i^{th} row equal to

$$D_2 f_i(\mathbf{x}(\mathbf{y}_2), \mathbf{y}^i),$$

it follows

$$|M(\mathbf{y}_1 - \mathbf{y}_2)| \leq \left(\sum_i K^2 |\mathbf{y}_1 - \mathbf{y}_2|^2 \right)^{1/2} = \sqrt{m} K |\mathbf{y}_1 - \mathbf{y}_2|. \quad (24.8)$$

Also, from 24.7,

$$J(\mathbf{x}^1, \dots, \mathbf{x}^n, \mathbf{y}_1)(\mathbf{x}(\mathbf{y}_1) - \mathbf{x}(\mathbf{y}_2)) = -M(\mathbf{y}_1 - \mathbf{y}_2) \quad (24.9)$$

and so from 24.8, 24.5,

$$\begin{aligned} |\mathbf{x}(\mathbf{y}_1) - \mathbf{x}(\mathbf{y}_2)| &= \left| J(\mathbf{x}^1, \dots, \mathbf{x}^n, \mathbf{y}_1)^{-1} M(\mathbf{y}_1 - \mathbf{y}_2) \right| \\ &= \left(\sum_{i=1}^n \left| e_i^T J(\mathbf{x}^1, \dots, \mathbf{x}^n, \mathbf{y}_1)^{-1} M(\mathbf{y}_1 - \mathbf{y}_2) \right|^2 \right)^{1/2} \\ &\leq \left(\sum_{i=1}^n K^2 |M(\mathbf{y}_1 - \mathbf{y}_2)|^2 \right)^{1/2} \leq \left(\sum_{i=1}^n K^2 (\sqrt{m} K |\mathbf{y}_1 - \mathbf{y}_2|)^2 \right)^{1/2} \\ &= K^2 \sqrt{mn} |\mathbf{y}_1 - \mathbf{y}_2| \end{aligned}$$

Thus $\mathbf{y} \rightarrow \mathbf{x}(\mathbf{y})$ is continuous near \mathbf{y}_0 .

Now let $\mathbf{y}_2 = \mathbf{y}, \mathbf{y}_1 = \mathbf{y} + h\mathbf{e}_k$ for small h . Then M depends on h and

$$\lim_{h \rightarrow 0} M(h) = D_2 \mathbf{f}(\mathbf{x}(\mathbf{y}), \mathbf{y})$$

thanks to the continuity of $\mathbf{y} \rightarrow \mathbf{x}(\mathbf{y})$ just shown. Also,

$$\frac{\mathbf{x}(\mathbf{y} + h\mathbf{e}_k) - \mathbf{x}(\mathbf{y})}{h} = -J(\mathbf{x}^1(h), \dots, \mathbf{x}^n(h), \mathbf{y} + h\mathbf{e}_k)^{-1} M(h) \mathbf{e}_k$$

Passing to a limit and using the formula for the inverse of a matrix in terms of the cofactor matrix, and the continuity of $\mathbf{y} \rightarrow \mathbf{x}(\mathbf{y})$ shown above, this yields

$$\frac{\partial \mathbf{x}}{\partial y_k} = -D_1 \mathbf{f}(\mathbf{x}(\mathbf{y}), \mathbf{y})^{-1} D_2 f_i(\mathbf{x}(\mathbf{y}), \mathbf{y}) \mathbf{e}_k$$

Then continuity of $\mathbf{y} \rightarrow \mathbf{x}(\mathbf{y})$ and the assumed continuity of the partial derivatives of \mathbf{f} shows that each partial derivative of $\mathbf{y} \rightarrow \mathbf{x}(\mathbf{y})$ exists and is continuous. ■

This implies the inverse function theorem given next.

Theorem 24.0.5 (inverse function theorem) Let $\mathbf{x}_0 \in U$, an open set in \mathbb{R}^n , and let $\mathbf{f} : U \rightarrow \mathbb{R}^n$. Suppose

$$\mathbf{f} \text{ is } C^1(U), \text{ and } D\mathbf{f}(\mathbf{x}_0)^{-1} \text{ exists.} \quad (24.10)$$

Then there exist open sets W , and V such that

$$\mathbf{x}_0 \in W \subseteq U, \quad (24.11)$$

$$\mathbf{f} : W \rightarrow V \text{ is one to one and onto,} \quad (24.12)$$

$$\mathbf{f}^{-1} \text{ is } C^1, \quad (24.13)$$

Proof: Apply the implicit function theorem to the function $\mathbf{F}(\mathbf{x}, \mathbf{y}) \equiv \mathbf{f}(\mathbf{x}) - \mathbf{y}$ where $\mathbf{y}_0 \equiv \mathbf{f}(\mathbf{x}_0)$. Thus the function $\mathbf{y} \rightarrow \mathbf{x}(\mathbf{y})$ defined in that theorem is \mathbf{f}^{-1} . Now let $W \equiv B(\mathbf{x}_0, \delta) \cap \mathbf{f}^{-1}(B(\mathbf{y}_0, \eta))$ and $V \equiv B(\mathbf{y}_0, \eta)$. This proves the theorem. ■

24.1 More Continuous Partial Derivatives

The implicit function theorem will now be improved slightly. If \mathbf{f} is C^k , it follows that the function which is implicitly defined is also C^k , not just C^1 , meaning all mixed partial derivatives of \mathbf{f} up to order k are continuous. Since the inverse function theorem comes as a case of the implicit function theorem, this shows that the inverse function also inherits the property of being C^k . First some notation is convenient. Let $\alpha = (\alpha_1, \dots, \alpha_n)$ where each α_i is a nonnegative integer. Then letting $|\alpha| = \sum_i \alpha_i$,

$$D^\alpha \mathbf{f}(\mathbf{x}) \equiv \frac{\partial^{|\alpha|} \mathbf{f}}{\partial \alpha_1 \partial \alpha_2 \dots \partial \alpha_n}(\mathbf{x}), \quad D^0 \mathbf{f}(\mathbf{x}) \equiv \mathbf{f}(\mathbf{x})$$

Theorem 24.1.1 (implicit function theorem) Suppose U is an open set in $\mathbb{F}^n \times \mathbb{F}^m$. Let $\mathbf{f} : U \rightarrow \mathbb{F}^m$ be in $C^k(U)$ and suppose

$$\mathbf{f}(\mathbf{x}_0, \mathbf{y}_0) = \mathbf{0}, \quad D_1 \mathbf{f}(\mathbf{x}_0, \mathbf{y}_0)^{-1} \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m). \quad (24.14)$$

Then there exist positive constants δ, η , such that for every $\mathbf{y} \in B(\mathbf{y}_0, \eta)$ there exists a unique $\mathbf{x}(\mathbf{y}) \in B(\mathbf{x}_0, \delta)$ such that

$$\mathbf{f}(\mathbf{x}(\mathbf{y}), \mathbf{y}) = \mathbf{0}. \quad (24.15)$$

Furthermore, the mapping $\mathbf{y} \rightarrow \mathbf{x}(\mathbf{y})$ is in $C^k(B(\mathbf{y}_0, \eta))$.

Proof: From the implicit function theorem $\mathbf{y} \rightarrow \mathbf{x}(\mathbf{y})$ is C^1 . It remains to show that it is C^k for $k > 1$ assuming that \mathbf{f} is C^k . From 24.15

$$\frac{\partial \mathbf{x}}{\partial y^i} = -D_1 \mathbf{f}(\mathbf{x}, \mathbf{y})^{-1} \frac{\partial \mathbf{f}}{\partial y^i}.$$

Thus the following formula holds for $q = 1$ and $|\alpha| = q$.

$$D^\alpha \mathbf{x}(\mathbf{y}) = \sum_{|\beta| \leq q} M_\beta(\mathbf{x}, \mathbf{y}) D^\beta \mathbf{f}(\mathbf{x}, \mathbf{y}) \quad (24.16)$$

where M_β is a matrix whose entries are differentiable functions of $D^\gamma \mathbf{x}$ for $|\gamma| < q$ and $D^\tau \mathbf{f}(\mathbf{x}, \mathbf{y})$ for $|\tau| \leq q$. This follows easily from the description of $D_1 \mathbf{f}(\mathbf{x}, \mathbf{y})^{-1}$ in terms of the cofactor matrix and the determinant of $D_1 \mathbf{f}(\mathbf{x}, \mathbf{y})$. Suppose 24.16 holds for $|\alpha| = q < k$. Then by induction, this yields \mathbf{x} is C^q . Then

$$\frac{\partial D^\alpha \mathbf{x}(\mathbf{y})}{\partial y^p} = \sum_{|\beta| \leq |\alpha|} \frac{\partial M_\beta(\mathbf{x}, \mathbf{y})}{\partial y^p} D^\beta \mathbf{f}(\mathbf{x}, \mathbf{y}) + M_\beta(\mathbf{x}, \mathbf{y}) \frac{\partial D^\beta \mathbf{f}(\mathbf{x}, \mathbf{y})}{\partial y^p}.$$

By the chain rule $\frac{\partial M_\beta(\mathbf{x}, \mathbf{y})}{\partial y^p}$ is a matrix whose entries are differentiable functions of

$$D^\tau \mathbf{f}(\mathbf{x}, \mathbf{y})$$

for $|\tau| \leq q+1$ and $D^\gamma \mathbf{x}$ for $|\gamma| < q+1$. It follows, since y^p was arbitrary, that for any $|\alpha| = q+1$, a formula like 24.16 holds with q being replaced by $q+1$. By induction, \mathbf{x} is C^k . ■

As a simple corollary, this yields the inverse function theorem. You just let $\mathbf{F}(\mathbf{x}, \mathbf{y}) = \mathbf{y} - \mathbf{f}(\mathbf{x})$ and apply the implicit function theorem.

Theorem 24.1.2 (inverse function theorem) *Let $\mathbf{x}_0 \in U \subseteq \mathbb{F}^n$ and let $\mathbf{f} : U \rightarrow \mathbb{F}^n$. Suppose for k a positive integer,*

$$\mathbf{f} \text{ is } C^k(U), \text{ and } D\mathbf{f}(\mathbf{x}_0)^{-1} \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^n). \quad (24.17)$$

Then there exist open sets W , and V such that

$$\mathbf{x}_0 \in W \subseteq U, \quad (24.18)$$

$$\mathbf{f} : W \rightarrow V \text{ is one to one and onto,} \quad (24.19)$$

$$\mathbf{f}^{-1} \text{ is } C^k. \quad (24.20)$$

24.2 The Method of Lagrange Multipliers

As an application of the implicit function theorem, consider the method of Lagrange multipliers. Recall the problem is to maximize or minimize a function subject to equality constraints. Let $f : U \rightarrow \mathbb{R}$ be a C^1 function where $U \subseteq \mathbb{R}^n$ and let

$$g_i(\mathbf{x}) = 0, \quad i = 1, \dots, m \quad (24.21)$$

be a collection of equality constraints with $m < n$. Now consider the system of nonlinear equations

$$\begin{aligned} f(\mathbf{x}) &= a \\ g_i(\mathbf{x}) &= 0, \quad i = 1, \dots, m. \end{aligned}$$

Recall \mathbf{x}_0 is a local maximum if $f(\mathbf{x}_0) \geq f(\mathbf{x})$ for all \mathbf{x} near \mathbf{x}_0 which also satisfies the constraints 24.21. A local minimum is defined similarly. Let $\mathbf{F} : U \times \mathbb{R} \rightarrow \mathbb{R}^{m+1}$ be defined by

$$\mathbf{F}(\mathbf{x}, a) \equiv \begin{pmatrix} f(\mathbf{x}) - a \\ g_1(\mathbf{x}) \\ \vdots \\ g_m(\mathbf{x}) \end{pmatrix}. \quad (24.22)$$

Now consider the $m+1 \times n$ matrix

$$\begin{pmatrix} f_{x_1}(\mathbf{x}_0) & \cdots & f_{x_n}(\mathbf{x}_0) \\ g_{1x_1}(\mathbf{x}_0) & \cdots & g_{1x_n}(\mathbf{x}_0) \\ \vdots & & \vdots \\ g_{mx_1}(\mathbf{x}_0) & \cdots & g_{mx_n}(\mathbf{x}_0) \end{pmatrix}.$$

If this matrix has rank $m+1$ then some $m+1 \times m+1$ submatrix has nonzero determinant. See Theorem 20.2.15. It follows from the implicit function theorem, there exists $m+1$ variables $x_{i_1}, \dots, x_{i_{m+1}}$ such that the system

$$\mathbf{F}(\mathbf{x}, a) = \mathbf{0} \quad (24.23)$$

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 (\mathbf{x}, a) to 24.23 for some \mathbf{x} close to \mathbf{x}_0 whenever a is in some open interval. Therefore, \mathbf{x}_0 cannot be either a local minimum or a local maximum. It follows that if \mathbf{x}_0 is either a local maximum or a local minimum, then the above matrix must have rank less than $m+1$. It follows that some row is a linear combination of the others. Thus there exist m scalars,

$$\lambda_1, \dots, \lambda_m,$$

and a scalar μ , not all zero such that

$$\mu \begin{pmatrix} f_{x_1}(\mathbf{x}_0) \\ \vdots \\ f_{x_n}(\mathbf{x}_0) \end{pmatrix} = \lambda_1 \begin{pmatrix} g_{1x_1}(\mathbf{x}_0) \\ \vdots \\ g_{1x_n}(\mathbf{x}_0) \end{pmatrix} + \cdots + \lambda_m \begin{pmatrix} g_{mx_1}(\mathbf{x}_0) \\ \vdots \\ g_{mx_n}(\mathbf{x}_0) \end{pmatrix}. \quad (24.24)$$

If the rank of the matrix

$$\begin{pmatrix} g_{1x_1}(\mathbf{x}_0) & \cdots & g_{mx_1}(\mathbf{x}_0) \\ \vdots & & \vdots \\ g_{1x_n}(\mathbf{x}_0) & \cdots & g_{mx_n}(\mathbf{x}_0) \end{pmatrix} \quad (24.25)$$

is m , then we can choose $\mu = 1$ because the columns span \mathbb{R}^m . Thus there are scalars λ_i such that

$$\begin{pmatrix} f_{x_1}(\mathbf{x}_0) \\ \vdots \\ f_{x_n}(\mathbf{x}_0) \end{pmatrix} = \lambda_1 \begin{pmatrix} g_{1x_1}(\mathbf{x}_0) \\ \vdots \\ g_{1x_n}(\mathbf{x}_0) \end{pmatrix} + \cdots + \lambda_m \begin{pmatrix} g_{mx_1}(\mathbf{x}_0) \\ \vdots \\ g_{mx_n}(\mathbf{x}_0) \end{pmatrix} \quad (24.26)$$

at every point \mathbf{x}_0 which is either a local maximum or a local minimum. This proves the following theorem.

Theorem 24.2.1 *Let U be an open subset of \mathbb{R}^n and let $f : U \rightarrow \mathbb{R}$ be a C^1 function. Then if $\mathbf{x}_0 \in U$ is either a local maximum or local minimum of f subject to the constraints 24.21, then 24.24 must hold for some scalars $\mu, \lambda_1, \dots, \lambda_m$ not all equal to zero. If the rank of the matrix in 24.25 is m , it follows 24.26 holds for some choice of the λ_i .*

Chapter 25

Line Integrals

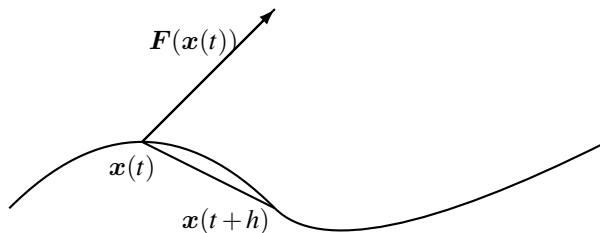
The concept of the integral can be extended to functions which are not defined on an interval of the real line but on some curve in \mathbb{R}^p . This is done by defining things in such a way that the more general concept reduces to the earlier notion. The arc length was discussed in the first part of this book which was on calculus of functions of one variable as was the notion of orientation of a curve.

25.1 Line Integrals and Work

Let C be a smooth curve contained in \mathbb{R}^p . A curve C is an “**oriented curve**” if the only parameterizations considered are those which lie in exactly one of the two equivalence classes, each of which is called an “**orientation**”. In simple language, orientation specifies a direction over which motion along the curve is to take place. Thus, it specifies the order in which the points of C are encountered. The pair of concepts consisting of the set of points making up the curve along with a direction of motion along the curve is called an **oriented curve**.

Definition 25.1.1 Suppose $\mathbf{F}(\mathbf{x}) \in \mathbb{R}^p$ is given for each $\mathbf{x} \in C$ where C is a smooth oriented curve and suppose $\mathbf{x} \rightarrow \mathbf{F}(\mathbf{x})$ is continuous. The mapping $\mathbf{x} \rightarrow \mathbf{F}(\mathbf{x})$ is called a **vector field**. In the case that $\mathbf{F}(\mathbf{x})$ is a force, it is called a **force field**.

Next the concept of work done by a force field \mathbf{F} on an object as it moves along the curve C , in the direction determined by the given orientation of the curve will be defined. This is new. Earlier the work done by a force which acts on an object moving in a straight line was discussed but here the object moves over a curve. In order to define what is meant by the work, consider the following picture.



In this picture, the work done by a constant force \mathbf{F} on an object which moves from the point $\mathbf{x}(t)$ to the point $\mathbf{x}(t+h)$ along the straight line shown would equal

$$\mathbf{F} \cdot (\mathbf{x}(t+h) - \mathbf{x}(t))$$

It is reasonable to assume this would be a good approximation to the work done in moving along the curve joining $\mathbf{x}(t)$ and $\mathbf{x}(t+h)$ provided h is small enough. Also, provided h is small,

$$\mathbf{x}(t+h) - \mathbf{x}(t) \approx \mathbf{x}'(t)h$$

where the wriggly equal sign indicates the two quantities are close. In the notation of Leibniz, one writes dt for h and

$$dW = \mathbf{F}(\mathbf{x}(t)) \cdot \mathbf{x}'(t) dt$$

or in other words,

$$\frac{dW}{dt} = \mathbf{F}(\mathbf{x}(t)) \cdot \mathbf{x}'(t).$$

Defining the total work done by the force at $t = 0$, corresponding to the first endpoint of the curve, to equal zero, the work would satisfy the following initial value problem.

$$\frac{dW}{dt} = \mathbf{F}(\mathbf{x}(t)) \cdot \mathbf{x}'(t), \quad W(a) = 0.$$

This motivates the following definition of work.

Definition 25.1.2 Let $\mathbf{F}(\mathbf{x})$ be given above. Then the work done by this force field on an object moving over the curve C in the direction determined by the specified orientation is defined as

$$\int_C \mathbf{F} \cdot d\mathbf{R} \equiv \int_a^b \mathbf{F}(\mathbf{x}(t)) \cdot \mathbf{x}'(t) dt$$

where the function \mathbf{x} is one of the allowed parameterizations of C in the given orientation of C . In other words, there is an interval $[a, b]$ and as t goes from a to b , $\mathbf{x}(t)$ moves in the direction determined from the given orientation of the curve.

Theorem 25.1.3 The symbol $\int_C \mathbf{F} \cdot d\mathbf{R}$, is well defined in the sense that every parametrization in the given orientation of C gives the same value for $\int_C \mathbf{F} \cdot d\mathbf{R}$.

Proof: Suppose $\mathbf{g} : [c, d] \rightarrow C$ is another allowed parametrization. Thus $\mathbf{g}^{-1} \circ \mathbf{f}$ is an increasing function ϕ . Then since ϕ is increasing, it follows from the change of variables formula that

$$\begin{aligned} \int_c^d \mathbf{F}(\mathbf{g}(s)) \cdot \mathbf{g}'(s) ds &= \int_a^b \mathbf{F}(\mathbf{g}(\phi(t))) \cdot \mathbf{g}'(\phi(t)) \phi'(t) dt \\ &= \int_a^b \mathbf{F}(\mathbf{f}(t)) \cdot \frac{d}{dt} (\mathbf{g}(\mathbf{g}^{-1} \circ \mathbf{f}(t))) dt = \int_a^b \mathbf{F}(\mathbf{f}(t)) \cdot \mathbf{f}'(t) dt. \blacksquare \end{aligned}$$

Regardless the physical interpretation of \mathbf{F} , this is called the **line integral**. When \mathbf{F} is interpreted as a force, the line integral measures the extent to which the motion over the curve in the indicated direction is aided by the force. If the net effect of the force on

the object is to impede rather than to aid the motion, this will show up as the work being negative.

Does the concept of work as defined here coincide with the earlier concept of work when the object moves over a straight line when acted on by a constant force? If it doesn't, then the above is not a good definition because it will contradict earlier and more basic constructions. Math is not like sectarian religions which are typically replete with inconsistencies and blatant contradictions.

Let \mathbf{p} and \mathbf{q} be two points in \mathbb{R}^n and suppose \mathbf{F} is a constant force acting on an object which moves from \mathbf{p} to \mathbf{q} along the straight line joining these points. Then the work done is $\mathbf{F} \cdot (\mathbf{q} - \mathbf{p})$. Is the same thing obtained from the above definition? Let $\mathbf{x}(t) \equiv \mathbf{p} + t(\mathbf{q} - \mathbf{p})$, $t \in [0, 1]$ be a parametrization for this oriented curve, the straight line in the direction from \mathbf{p} to \mathbf{q} . Then $\mathbf{x}'(t) = \mathbf{q} - \mathbf{p}$ and $\mathbf{F}(\mathbf{x}(t)) = \mathbf{F}$. Therefore, the above definition yields $\int_0^1 \mathbf{F} \cdot (\mathbf{q} - \mathbf{p}) dt = \mathbf{F} \cdot (\mathbf{q} - \mathbf{p})$. Therefore, the new definition adds to but does not contradict the old one. Therefore, it is not unreasonable to use this as the definition.

Example 25.1.4 Suppose for $t \in [0, \pi]$ the position of an object is given by $\mathbf{r}(t) = t\mathbf{i} + \cos(2t)\mathbf{j} + \sin(2t)\mathbf{k}$. Also suppose there is a force field defined on \mathbb{R}^3 , $\mathbf{F}(x, y, z) \equiv 2xy\mathbf{i} + x^2\mathbf{j} + \mathbf{k}$. Find $\int_C \mathbf{F} \cdot d\mathbf{R}$ where C is the curve traced out by this object which has the orientation determined by the direction of increasing t .

To find this line integral use the above definition and write

$$\int_C \mathbf{F} \cdot d\mathbf{R} = \int_0^\pi (2t(\cos(2t)), t^2, 1) \cdot (1, -2\sin(2t), 2\cos(2t)) dt$$

In evaluating this replace the x in the formula for \mathbf{F} with t , the y in the formula for \mathbf{F} with $\cos(2t)$ and the z in the formula for \mathbf{F} with $\sin(2t)$ because these are the values of these variables which correspond to the value of t . Taking the dot product, this equals the following integral.

$$\int_0^\pi (2t \cos 2t - 2(\sin 2t)t^2 + 2 \cos 2t) dt = \pi^2$$

Example 25.1.5 Let C denote the oriented curve obtained by $\mathbf{r}(t) = (t, \sin t, t^3)$ where the orientation is determined by increasing t for $t \in [0, 2]$. Also let

$$\mathbf{F} = (x, y, xz + z)$$

Find $\int_C \mathbf{F} \cdot d\mathbf{R}$.

You use the definition.

$$\begin{aligned} \int_C \mathbf{F} \cdot d\mathbf{R} &= \int_0^2 (t, \sin(t), (t+1)t^3) \cdot (1, \cos(t), 3t^2) dt \\ &= \int_0^2 (t + \sin(t)\cos(t) + 3(t+1)t^5) dt = \frac{1251}{14} - \frac{1}{2}\cos^2(2). \end{aligned}$$

Suppose you have a curve specified by $\mathbf{r}(s) = (x(s), y(s), z(s))$ and it has the property that $|\mathbf{r}'(s)| = 1$ for all $s \in [0, b]$. Then the length of this curve for s between 0 and s_1 is $\int_0^{s_1} |\mathbf{r}'(s)| ds = \int_0^{s_1} 1 ds = s_1$. This parameter is therefore called arc length because the length of the curve up to s equals s . Now you can always change the parameter to be arc length.

Proposition 25.1.6 Suppose C is an oriented smooth curve parameterized by $\mathbf{r}(t)$ for $t \in [a, b]$. Then letting l denote the total length of C , there exists $\mathbf{R}(s)$, $s \in [0, l]$ another parametrization for this curve which preserves the orientation and such that $|\mathbf{R}'(s)| = 1$ so that s is arc length.

Prove: Let $\phi(t) \equiv \int_a^t |\mathbf{r}'(\tau)| d\tau \equiv s$. Then s is an increasing function of t because

$$\frac{ds}{dt} = \phi'(t) = |\mathbf{r}'(t)| > 0.$$

Now define $\mathbf{R}(s) \equiv \mathbf{r}(\phi^{-1}(s))$. Then

$$\mathbf{R}'(s) = \mathbf{r}'(\phi^{-1}(s)) (\phi^{-1})'(s) = \frac{\mathbf{r}'(\phi^{-1}(s))}{|\mathbf{r}'(\phi^{-1}(s))|}$$

and so $|\mathbf{R}'(s)| = 1$ as claimed. $\mathbf{R}(l) = \mathbf{r}(\phi^{-1}(l)) = \mathbf{r}(\phi^{-1}(\int_a^b |\mathbf{r}'(\tau)| d\tau)) = \mathbf{r}(b)$ and $\mathbf{R}(0) = \mathbf{r}(\phi^{-1}(0)) = \mathbf{r}(a)$ and \mathbf{R} delivers the same set of points in the same order as \mathbf{r} because $\frac{ds}{dt} > 0$. ■

The arc length parameter is just like any other parameter, in so far as considerations of line integrals are concerned, because it was shown above that line integrals are independent of parametrization. However, when things are defined in terms of the arc length parametrization, it is clear they depend only on geometric properties of the curve itself and for this reason, the arc length parametrization is important in differential geometry.

Definition 25.1.7 Recall piecewise smooth curves are just smooth curves joined together at a succession of points $\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_m$. If C is such a curve which goes from \mathbf{p}_1 then to \mathbf{p}_2 then to \mathbf{p}_3 etc. one defines

$$\int_C \mathbf{F} \cdot d\mathbf{R} \equiv \int_{C_{\mathbf{p}_1 \mathbf{p}_2}} \mathbf{F} \cdot d\mathbf{R} + \int_{C_{\mathbf{p}_2 \mathbf{p}_3}} \mathbf{F} \cdot d\mathbf{R} + \dots + \int_{C_{\mathbf{p}_{(n-1)n}}} \mathbf{F} \cdot d\mathbf{R}$$

25.2 Conservative Fields and Notation

Conservative vector fields are the gradient of some scalar function.

Proposition 25.2.1 Suppose C is a piecewise smooth curve which goes from \mathbf{p} to \mathbf{q} . Also suppose that $\mathbf{F}(\mathbf{x}) = \nabla \phi(\mathbf{x})$. Then $\int_C \mathbf{F} \cdot d\mathbf{R} = \phi(\mathbf{q}) - \phi(\mathbf{p})$.

Proof: Say $\mathbf{r}(t)$, $t \in [a_i, b_i]$ is a parametrization for C going from \mathbf{x}_{i-1} to \mathbf{x}_i and \mathbf{r} is a parameterization for the smooth curve from \mathbf{x}_{i-1} to \mathbf{x}_i with $\mathbf{x}_0 = \mathbf{p}$ and $\mathbf{x}_m = \mathbf{q}$. Then, from the chain rule,

$$\begin{aligned} \int_C \mathbf{F} \cdot d\mathbf{R} &= \sum_{i=1}^m \int_{a_i}^{b_i} \nabla \phi(\mathbf{r}_i(t)) \cdot \mathbf{r}'_i(t) dt = \sum_{i=1}^m \int_{a_i}^{b_i} \frac{d}{dt} (\phi(\mathbf{r}_i(t))) dt \\ &= \sum_{i=1}^m \phi(\mathbf{x}_i) - \phi(\mathbf{x}_{i-1}) = \phi(\mathbf{q}) - \phi(\mathbf{p}) \quad \blacksquare \end{aligned}$$

Note how this says that the integral is path independent, depending only on the values of the function ϕ , called a potential function, at the end points.

Definition 25.2.2 Let $\mathbf{F}(x, y, z) = (P(x, y, z), Q(x, y, z), R(x, y, z))$ and let C be an oriented curve. Then another way to write $\int_C \mathbf{F} \cdot d\mathbf{R}$ is $\int_C Pdx + Qdy + Rdz$

This last is referred to as the integral of a **differential form**, $Pdx + Qdy + Rdz$. The study of differential forms is important. Formally, $d\mathbf{R} = (dx, dy, dz)$ and so the integrand in the above is formally $\mathbf{F} \cdot d\mathbf{R}$. Other occurrences of this notation are handled similarly in 2 or higher dimensions.

25.3 Exercises

1. Let $\mathbf{r}(t) = (\ln(t), \frac{t^2}{2}, \sqrt{2}t)$ for $t \in [1, 2]$. Find the length of this curve.
2. Let $\mathbf{r}(t) = (\frac{2}{3}t^{3/2}, t, t)$ for $t \in [0, 1]$. Find the length of this curve.
3. Let $\mathbf{r}(t) = (t, \cos(3t), \sin(3t))$ for $t \in [0, 1]$. Find the length of this curve.
4. Suppose for $t \in [0, \pi]$ the position of an object is given by $\mathbf{r}(t) = t\mathbf{i} + \cos(2t)\mathbf{j} + \sin(2t)\mathbf{k}$. Also suppose there is a force field defined on \mathbb{R}^3 ,

$$\mathbf{F}(x, y, z) \equiv 2xy\mathbf{i} + (x^2 + 2zy)\mathbf{j} + y^2\mathbf{k}$$

Find the work

$$\int_C \mathbf{F} \cdot d\mathbf{R}$$

where C is the curve traced out by this object having the orientation determined by the direction of increasing t .

5. In the following, a force field is specified followed by the parametrization of a curve. Find the work.
 - (a) $\mathbf{F} = (x, y, z), \mathbf{r}(t) = (t, t^2, t + 1), t \in [0, 1]$
 - (b) $\mathbf{F} = (x - y, y + z, z), \mathbf{r}(t) = (\cos(t), t, \sin(t)), t \in [0, \pi]$
 - (c) $\mathbf{F} = (x^2, y^2, z + x), \mathbf{r}(t) = (t, 2t, t + t^2), t \in [0, 1]$
 - (d) $\mathbf{F} = (z, y, x), \mathbf{r}(t) = (t^2, 2t, t), t \in [0, 1]$
6. The curve consists of straight line segments which go from $(0, 0, 0)$ to $(1, 1, 1)$ and finally to $(1, 2, 3)$. Find the work done if the force field is
 - (a) $\mathbf{F} = (2xy, x^2 + 2y, 1)$
 - (b) $\mathbf{F} = (yz^2, xz^2, 2xyz + 1)$
 - (c) $\mathbf{F} = (\cos x, -\sin y, 1)$
 - (d) $\mathbf{F} = (2x \sin y, x^2 \cos y, 1)$

7. Show the vector fields in the preceding problems are respectively

$$\nabla(x^2y + y^2 + z), \nabla(xyz^2 + z), \nabla(\sin x + \cos y + z - 1)$$

and $\nabla(x^2 \sin y + z)$. Thus each of these vector fields is of the form ∇f where f is a function of three variables. Use Proposition 25.2.1 to evaluate each of the line integrals. Compare with what you get by doing it directly.

8. Suppose for $t \in [0, 1]$ the position of an object is given by $\mathbf{r}(t) = t\mathbf{i} + t\mathbf{j} + t\mathbf{k}$. Also suppose there is a force field defined on \mathbb{R}^3 , $\mathbf{F}(x, y, z) \equiv yz\mathbf{i} + xz\mathbf{j} + xy\mathbf{k}$. Find $\int_C \mathbf{F} \cdot d\mathbf{R}$ where C is the curve traced out by this object which has the orientation determined by the direction of increasing t . Repeat the problem for $\mathbf{r}(t) = t\mathbf{i} + t^2\mathbf{j} + t\mathbf{k}$. Verify a scalar potential is $\phi(x, y, z) = xyz$.
9. Here is a vector field $(y, x + z^2, 2yz)$ and here is the parametrization of a curve C . $\mathbf{R}(t) = (\cos 2t, 2\sin 2t, t)$ where t goes from 0 to $\pi/4$. Find $\int_C \mathbf{F} \cdot d\mathbf{R}$.
10. If f and g are both increasing functions, show that $f \circ g$ is an increasing function also. Assume anything you like about the domains of the functions.
11. Suppose for $t \in [0, 3]$ the position of an object is given by $\mathbf{r}(t) = t\mathbf{i} + t\mathbf{j} + t\mathbf{k}$. Also suppose there is a force field defined on \mathbb{R}^3 , $\mathbf{F}(x, y, z) \equiv yz\mathbf{i} + xz\mathbf{j} + xy\mathbf{k}$. Find $\int_C \mathbf{F} \cdot d\mathbf{R}$ where C is the curve traced out by this object which has the orientation determined by the direction of increasing t . Repeat the problem for $\mathbf{r}(t) = t\mathbf{i} + t^2\mathbf{j} + t\mathbf{k}$.
12. Suppose for $t \in [0, 1]$ the position of an object is given by $\mathbf{r}(t) = t\mathbf{i} + t\mathbf{j} + t\mathbf{k}$. Also suppose there is a force field defined on \mathbb{R}^3 , $\mathbf{F}(x, y, z) \equiv z\mathbf{i} + xz\mathbf{j} + xy\mathbf{k}$. Find $\int_C \mathbf{F} \cdot d\mathbf{R}$ where C is the curve traced out by this object which has the orientation determined by the direction of increasing t . Repeat the problem for $\mathbf{r}(t) = t\mathbf{i} + t^2\mathbf{j} + t\mathbf{k}$.
13. Let $\mathbf{F}(x, y, z)$ be a given force field and suppose it acts on an object having mass m on a curve with parametrization, $(x(t), y(t), z(t))$ for $t \in [a, b]$. Show directly that the work done equals the difference in the kinetic energy. **Hint:**

$$\begin{aligned} & \int_a^b \mathbf{F}(x(t), y(t), z(t)) \cdot (x'(t), y'(t), z'(t)) dt \\ &= \int_a^b m(x''(t), y''(t), z''(t)) \cdot (x'(t), y'(t), z'(t)) dt, \end{aligned}$$

14. Suppose for $t \in [0, 2\pi]$ the position of an object is given by

$$\mathbf{r}(t) = 2t\mathbf{i} + \cos(t)\mathbf{j} + \sin(t)\mathbf{k}.$$

Also suppose there is a force field defined on \mathbb{R}^3 ,

$$\mathbf{F}(x, y, z) \equiv 2xy\mathbf{i} + (x^2 + 2zy)\mathbf{j} + y^2\mathbf{k}.$$

Find the work $\int_C \mathbf{F} \cdot d\mathbf{R}$ where C is the curve traced out by this object which has the orientation determined by the direction of increasing t .

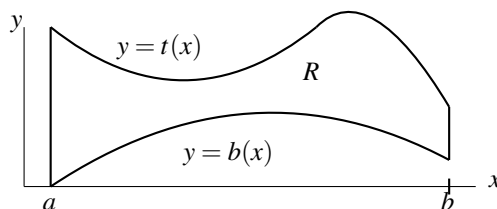
15. Here is a vector field $(y, x^2 + z, 2yz)$ and here is the parametrization of a curve C . $\mathbf{R}(t) = (\cos 2t, 2\sin 2t, t)$ where t goes from 0 to $\pi/4$. Find $\int_C \mathbf{F} \cdot d\mathbf{R}$.

Chapter 26

The Riemannnn Integral on \mathbb{R}^p

26.1 Methods for Double Integrals

This chapter is on the Riemannnn integral for a function of p variables. It begins by introducing the basic concepts and applications of the integral. The general considerations including the definition of the integral and proofs of theorems are left till later. These are very difficult topics and are likely better considered in the context of the Lebesgue integral. Consider the following region which is labeled R .



We will consider the following iterated integral which makes sense for any continuous function $f(x, y)$.

$$\int_a^b \int_{b(x)}^{t(x)} f(x, y) dy dx$$

It means just exactly what the notation suggests it does. You fix x and then you do the inside integral

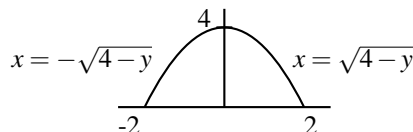
$$\int_{b(x)}^{t(x)} f(x, y) dy$$

This yields a function of x which will end up being continuous. You then do $\int_a^b dx$ to this continuous function.

What was it about the above region which made it possible to set up such an iterated integral? It was just this: You have a curve on the top $y = t(x)$, and a curve on the bottom $y = b(x)$ for $x \in [a, b]$. You could have set up a similar iterated integral if you had a region in which there was a curve on the left and a curve on the right for y in some interval. Here is an example.

Example 26.1.1 Suppose $t(x) = 4 - x^2$, $b(x) = 0$ and $a = -2, b = 2$. Compute the iterated integral described above for $f(x, y) = xy + y$.

Consider the graphs of these functions.



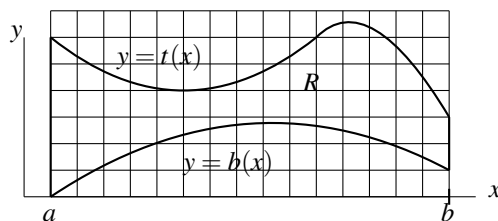
Filling in the limits as above, we obtain

$$\int_{-2}^2 \int_0^{4-x^2} (xy+y) dy dx = \int_{-2}^2 \frac{1}{2} (x^2-4)^2 (x+1) dx = \frac{256}{15}$$

Of course one could do the iterated integral in the other order for this example. In this case, you would be considering a curve on the left $x = -\sqrt{4-y}$, a curve on the right $x = \sqrt{4-y}$, and $y \in [0, 4]$. Thus this iterated integral would be of the form

$$\int_0^4 \int_{-\sqrt{4-y}}^{\sqrt{4-y}} (xy+y) dx dy = \int_0^4 2y\sqrt{4-y} dy = \frac{256}{15}$$

Why should it be the case that these two iterated integrals are equal? This involves a consideration of what you are computing when you do such an iterated integral. First note that in the general example given above involving $t(x)$, $b(x)$, it would not have been at all convenient to have done the iterated integral in the other order. So what is it you are getting? Consider the first illustration where the region is between $y = b(x)$ and $y = t(x)$. Consider the following picture



For simplicity, we let the distance between the vertical lines be Δx and the distance between the horizontal lines be Δy . We will only consider those rectangles which intersect the region R . Thus we will have $a = x_0 < x_1 < \dots < x_n = b$ and in the vertical direction, we will have

$$y_{im(i)} < y_{i(m(i)+1)} < \dots < y_{iM(i)}$$

where $m(i)$ is the largest such that $y_{im(i)}$ is no larger than $b(x_i)$ and $M(i)$ is the smallest such that $y_{iM(i)}$ is as large as $y(x_i)$. Then the iterated integral should satisfy the following approximate equalities

$$\begin{aligned} \int_a^b \int_{b(x)}^{t(x)} f(x,y) dy dx &= \sum_{i=1}^n \int_{x_{i-1}}^{x_i} \int_{b(x)}^{t(x)} f(x,y) dy dx \\ &\approx \sum_{i=1}^n \int_{x_{i-1}}^{x_i} \int_{b(x_i)}^{t(x_i)} f(x_i,y) dy dx \\ &\approx \sum_{i=1}^n \int_{x_{i-1}}^{x_i} \sum_{j=m(i)}^{M(i)} f(x_i, y_{ij}) \Delta y dx \\ &= \sum_{i=1}^n \sum_{j=m(i)}^{M(i)} f(x_i, y_{ij}) \Delta y \Delta x \end{aligned}$$

where we can extend f to be 0 off the region R . We would expect these approximations to improve as $\Delta x, \Delta y$ converge to 0, provided that the boundary of R is sufficiently “thin”. Thus the iterated integral ought to equal the number to which the “Riemann sums” represented by the last expression converge as $\Delta x, \Delta y \rightarrow 0$. That sum on the right is really just a systematic way of taking the value of the function at a point of a rectangle which intersects R , multiplying by the area of the rectangle containing this point and adding them together. It would have worked out similarly if we had been able to do the iterated integral in the other order, provided the boundary of R is “thin” enough, a completely stupid consideration which is not needed in the context of the Lebesgue integral. We would still have a sum of values of the function times areas of little rectangles. This is why it is entirely reasonable to expect the iterated integrals in two different orders to be equal. It is also why the iterated integral is approximating something which we call the Riemann integral.

For another more precise explanation for equality of iterated integrals in the case where the function is continuous, see Problem 7 on Page 248. For the whole story, see the chapter on the Lebesgue integral.

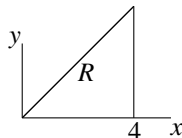
Definition 26.1.2 Let R be a bounded region in the xy plane and let f be a bounded function defined on R . We say f is Riemann integrable if there exists a number, denoted by $\int_R f dA$ and called the Riemann integral such that if $\varepsilon > 0$ is given, then whenever one imposes a sufficiently fine mesh enclosing R and considers the finitely many rectangles which intersect R , numbered as $\{Q_i\}_{i=1}^m$ and a point $(x_i, y_i) \in Q_i$, it follows that

$$\left| \int_R f dA - \sum_i f(x_i, y_i) \text{area}(Q_i) \right| < \varepsilon$$

It is $\int_R f dA$ which is of interest. The iterated integral should always be considered as a tool for computing this number. When this is kept in mind, things become less confusing. Also, it is helpful to consider $\int_R f dA$ as a kind of a glorified sum. It means to take the value of f at a point and multiply by a little chunk of area dA and then add these together, hence the integral sign which is really just an elongated symbol for a sum.

The careful explanation of these ideas is contained later in a special chapter devoted to the theory of the integral. I have presented there the Lebesgue integral because it is much easier to understand and use although it is more abstract.

Example 26.1.3 Let $f(x, y) = x^2y + yx$ for $(x, y) \in R$ where R is the triangular region defined to be in the first quadrant, below the line $y = x$ and to the left of the line $x = 4$. Find $\int_R f dA$.



From the above discussion,

$$\int_R f dA = \int_0^4 \int_0^x (x^2y + yx) dy dx$$

The reason for this is that x goes from 0 to 4 and for each fixed x between 0 and 4, y goes from 0 to the slanted line, $y = x$, the function being defined to be 0 for larger y . Thus y goes

from 0 to x . This explains the inside integral. Now $\int_0^x (x^2y + yx) dy = \frac{1}{2}x^4 + \frac{1}{2}x^3$ and so

$$\int_R f dA = \int_0^4 \left(\frac{1}{2}x^4 + \frac{1}{2}x^3 \right) dx = \frac{672}{5}.$$

What of integration in a different order? Lets put the integral with respect to y on the outside and the integral with respect to x on the inside. Then

$$\int_R f dA = \int_0^4 \int_y^4 (x^2y + yx) dx dy$$

For each y between 0 and 4, the variable x , goes from y to 4.

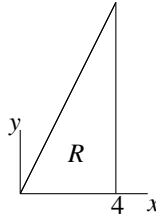
$$\int_y^4 (x^2y + yx) dx = \frac{88}{3}y - \frac{1}{3}y^4 - \frac{1}{2}y^3$$

Now

$$\int_R f dA = \int_0^4 \left(\frac{88}{3}y - \frac{1}{3}y^4 - \frac{1}{2}y^3 \right) dy = \frac{672}{5}.$$

Here is a similar example.

Example 26.1.4 Let $f(x, y) = x^2y$ for $(x, y) \in R$ where R is the triangular region defined to be in the first quadrant, below the line $y = 2x$ and to the left of the line $x = 4$. Find $\int_R f dA$.



Put the integral with respect to x on the outside first. Then

$$\int_R f dA = \int_0^4 \int_0^{2x} (x^2y) dy dx$$

because for each $x \in [0, 4]$, y goes from 0 to $2x$. Then

$$\int_0^{2x} (x^2y) dy = 2x^4$$

and so

$$\int_R f dA = \int_0^4 (2x^4) dx = \frac{2048}{5}$$

Now do the integral in the other order. Here the integral with respect to y will be on the outside. What are the limits of this integral? Look at the triangle and note that x goes from 0 to 4 and so $2x = y$ goes from 0 to 8. Now for fixed y between 0 and 8, where does x go? It goes from the x coordinate on the line $y = 2x$ which corresponds to this y to 4. What is the x coordinate on this line which goes with y ? It is $x = y/2$. Therefore, the iterated integral is

$$\int_0^8 \int_{y/2}^4 (x^2y) dx dy.$$

Now

$$\int_{y/2}^4 (x^2 y) dx = \frac{64}{3} y - \frac{1}{24} y^4$$

and so

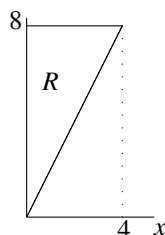
$$\int_R f dA = \int_0^8 \left(\frac{64}{3} y - \frac{1}{24} y^4 \right) dy = \frac{2048}{5}$$

the same answer.

A few observations are in order here. In finding $\int_S f dA$ there is no problem in setting things up if S is a rectangle. However, if S is not a rectangle, the procedure **always** is agonizing. A good rule of thumb is that if what you do is easy it will be wrong. There are no shortcuts! There are no quick fixes which require no thought! Pain and suffering is inevitable and you must not expect it to be otherwise. Always draw a picture and then begin **agonizing** over the correct limits. Even when you are careful you will make lots of mistakes until you get used to the process.

Sometimes an integral can be evaluated in one order but not in another.

Example 26.1.5 For R as shown below, find $\int_R \sin(y^2) dA$.



Setting this up to have the integral with respect to y on the inside yields

$$\int_0^4 \int_{2x}^8 \sin(y^2) dy dx.$$

Unfortunately, there is no antiderivative in terms of elementary functions for $\sin(y^2)$ so there is an immediate problem in evaluating the inside integral. It doesn't work out so the next step is to do the integration in another order and see if some progress can be made. This yields

$$\int_0^8 \int_0^{y/2} \sin(y^2) dx dy = \int_0^8 \frac{y}{2} \sin(y^2) dy$$

and $\int_0^8 \frac{y}{2} \sin(y^2) dy = -\frac{1}{4} \cos 64 + \frac{1}{4}$ which you can verify by making the substitution, $u = y^2$. Thus

$$\int_R \sin(y^2) dy = -\frac{1}{4} \cos 64 + \frac{1}{4}.$$

This illustrates an important idea. The integral $\int_R \sin(y^2) dA$ is defined as a number. It is the unique number between all the upper sums and all the lower sums. Finding it is another matter. In this case it was possible to find it using one order of integration but not the other. The iterated integral in this other order also is defined as a number but it cannot be found directly without interchanging the order of integration. Of course sometimes nothing you try will work out.

26.1.1 Density and Mass

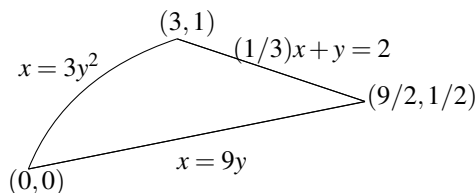
Consider a two dimensional material. Of course there is no such thing but a flat plate might be modeled as one. The density ρ is a function of position and is defined as follows. Consider a small chunk of area dA located at the point whose Cartesian coordinates are (x, y) . Then the mass of this small chunk of material is given by $\rho(x, y) dA$. Thus if the material occupies a region in two dimensional space U , the total mass of this material would be

$$\int_U \rho dA$$

In other words you integrate the density to get the mass. Now by letting ρ depend on position, you can include the case where the material is not homogeneous. Here is an example.

Example 26.1.6 Let $\rho(x, y)$ denote the density of the plane region determined by the curves $\frac{1}{3}x + y = 2$, $x = 3y^2$, and $x = 9y$. Find the total mass if $\rho(x, y) = y$.

You need to first draw a picture of the region R . A rough sketch follows.



This region is in two pieces, one having the graph of $x = 9y$ on the bottom and the graph of $x = 3y^2$ on the top and another piece having the graph of $x = 9y$ on the bottom and the graph of $\frac{1}{3}x + y = 2$ on the top. Therefore, in setting up the integrals, with the integral with respect to x on the outside, the double integral equals the following sum of iterated integrals.

$$\overbrace{\int_0^3 \int_{x/9}^{\sqrt{x/3}} y dy dx}^{\text{has } x=3y^2 \text{ on top}} + \overbrace{\int_3^{9/2} \int_{x/9}^{2-\frac{1}{3}x} y dy dx}^{\text{has } \frac{1}{3}x+y=2 \text{ on top}}$$

You notice it is not necessary to have a perfect picture, just one which is good enough to figure out what the limits should be. The dividing line between the two cases is $x = 3$ and this was shown in the picture. Now it is only a matter of evaluating the iterated integrals which in this case is routine and gives 1.

26.2 Exercises

1. Evaluate the iterated integral and then write the iterated integral with the order of integration reversed. $\int_0^4 \int_0^{3y} x dx dy$.
2. Evaluate the iterated integral and then write the iterated integral with the order of integration reversed. $\int_0^3 \int_0^{3y} y dx dy$.

3. Evaluate the iterated integral and then write the iterated integral with the order of integration reversed. $\int_0^2 \int_0^{2y} (x+1) dx dy$.
4. Evaluate the iterated integral and then write the iterated integral with the order of integration reversed. $\int_0^3 \int_0^y \sin(x) dx dy$.
5. Evaluate the iterated integral and then write the iterated integral with the order of integration reversed. $\int_0^1 \int_0^y \exp(y) dx dy$.
6. Let $\rho(x, y)$ denote the density of the plane region closest to $(0, 0)$ which is between the curves $x+2y=3$, $x=y^2$, and $x=0$. Find the total mass if $\rho(x, y) = y$. Set up the integral in terms of $dx dy$ and in terms of $dy dx$.
7. Let $\rho(x, y)$ denote the density of the plane region determined by the curves $x+2y=3$, $x=y^2$, and $x=4y$. Find the total mass if $\rho(x, y) = x$. Set up the integral in terms of $dx dy$ and $dy dx$.
8. Let $\rho(x, y)$ denote the density of the plane region determined by the curves $y=2x$, $y=x$, $x+y=3$. Find the total mass if $\rho(x, y) = y+1$. Set up the integrals in terms of $dx dy$ and $dy dx$.
9. Let $\rho(x, y)$ denote the density of the plane region determined by the curves $y=3x$, $y=x$, $2x+y=4$. Find the total mass if $\rho(x, y) = 1$.
10. Let $\rho(x, y)$ denote the density of the plane region determined by the curves $y=3x$, $y=x$, $x+y=2$. Find the total mass if $\rho(x, y) = x+1$. Set up the integrals in terms of $dx dy$ and $dy dx$.
11. Let $\rho(x, y)$ denote the density of the plane region determined by the curves $y=5x$, $y=x$, $5x+2y=10$. Find the total mass if $\rho(x, y) = 1$. Set up the integrals in terms of $dx dy$ and $dy dx$.
12. Find $\int_0^4 \int_{y/2}^2 \frac{1}{x} e^{2\frac{y}{x}} dx dy$. You might need to interchange the order of integration.
13. Find $\int_0^8 \int_{y/2}^4 \frac{1}{x} e^{3\frac{y}{x}} dx dy$.
14. Find $\int_0^{\frac{1}{3}\pi} \int_x^{\frac{1}{3}\pi} \frac{\sin y}{y} dy dx$.
15. Find $\int_0^{\frac{1}{2}\pi} \int_x^{\frac{1}{2}\pi} \frac{\sin y}{y} dy dx$.
16. Find $\int_0^\pi \int_x^\pi \frac{\sin y}{y} dy dx$.
17. * Evaluate the iterated integral and then write the iterated integral with the order of integration reversed. $\int_{-3}^3 \int_{-x}^x x^2 dy dx$

You should get

$$\int_3^0 \int_{-3}^{-y} x^2 dx dy + \int_0^{-3} \int_{-3}^y x^2 dx dy + \int_0^3 \int_y^3 x^2 dx dy + \int_{-3}^0 \int_{-y}^3 x^2 dx dy$$

This is a very interesting example which shows that iterated integrals have a life of their own, not just as a method for evaluating double integrals.

18. * Evaluate the iterated integral and then write the iterated integral with the order of integration reversed. $\int_{-2}^2 \int_{-x}^x x^2 dy dx$.

26.3 Methods for Triple Integrals

26.3.1 Definition of the Integral

The integral of a function of three variables is similar to the integral of a function of two variables. In this case, the term: “mesh” refers to a collection of little boxes which covers a given region in R .

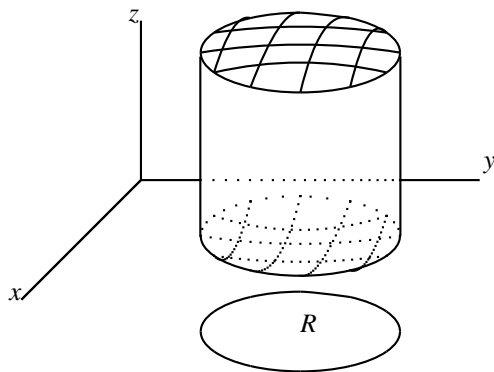
Definition 26.3.1 *Let R be a bounded region in the \mathbb{R}^3 and let f be a bounded function defined on R . We say f is Riemannn integrable if there exists a number, denoted by $\int_R f dV$ and called the Riemannn integral such that if $\epsilon > 0$ is given, then whenever one imposes a sufficiently fine mesh enclosing R and considers the finitely many boxes which intersect R , numbered as $\{Q_i\}_{i=1}^m$ and a point $(x_i, y_i, z_i) \in Q_i$, it follows that*

$$\left| \int_R f dV - \sum_i f(x_i, y_i, z_i) \text{volume}(Q_i) \right| < \epsilon$$

Of course one can continue generalizing to higher dimensions by analogy. By exactly similar reasoning to the case of integrals of functions of two variables, we can consider iterated integrals as a tool for finding the Riemannn integral of a function of three or more variables.

26.3.2 Iterated Integrals

As before, the integral is often computed by using an iterated integral. In general it is impossible to set up an iterated integral for finding $\int_E f dV$ for arbitrary regions, E but when the region is sufficiently simple, one can make progress. Suppose the region E over which the integral is to be taken is of the form $E = \{(x, y, z) : a(x, y) \leq z \leq b(x, y)\}$ for $(x, y) \in R$, a two dimensional region. This is illustrated in the following picture in which the bottom surface is the graph of $z = a(x, y)$ and the top is the graph of $z = b(x, y)$.



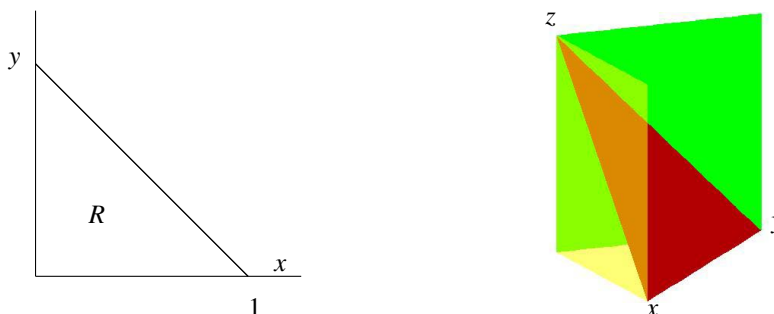
Then

$$\int_E f dV = \int_R \int_{a(x,y)}^{b(x,y)} f(x, y, z) dz dA$$

It might be helpful to think of $dV = dzdA$. Now $\int_{a(x,y)}^{b(x,y)} f(x,y,z) dz$ is a function of x and y and so you have reduced the triple integral to a double integral over R of this function of x and y . Similar reasoning would apply if the region in \mathbb{R}^3 were of the form $\{(x,y,z) : a(y,z) \leq x \leq b(y,z)\}$ or $\{(x,y,z) : a(x,z) \leq y \leq b(x,z)\}$.

Example 26.3.2 Find the volume of the region E in the first octant between $z = 1 - (x + y)$ and $z = 0$.

In this case, R is the region shown.



Thus the region E is between the plane $z = 1 - (x + y)$ on the top, $z = 0$ on the bottom, and over R shown above. Thus

$$\int_E 1 dV = \int_R \int_0^{1-(x+y)} dz dA = \int_0^1 \int_0^{1-x} \int_0^{1-(x+y)} dz dy dx = \frac{1}{6}$$

Of course iterated integrals have a life of their own although this will not be explored here. You can just write them down and go to work on them. Here are some examples.

Example 26.3.3 Find $\int_2^3 \int_3^x \int_{3y}^x (x - y) dz dy dx$.

The inside integral yields $\int_{3y}^x (x - y) dz = x^2 - 4xy + 3y^2$. Next this must be integrated with respect to y to give $\int_3^x (x^2 - 4xy + 3y^2) dy = -3x^2 + 18x - 27$. Finally the third integral gives

$$\int_2^3 \int_3^x \int_{3y}^x (x - y) dz dy dx = \int_2^3 (-3x^2 + 18x - 27) dx = -1.$$

Example 26.3.4 Find $\int_0^\pi \int_0^{3y} \int_0^{y+z} \cos(x + y) dx dz dy$.

The inside integral is $\int_0^{y+z} \cos(x + y) dx = 2 \cos z \sin y \cos y + 2 \sin z \cos^2 y - \sin z - \sin y$. Now this has to be integrated.

$$\begin{aligned} & \int_0^{3y} \int_0^{y+z} \cos(x + y) dx dz \\ &= \int_0^{3y} (2 \cos z \sin y \cos y + 2 \sin z \cos^2 y - \sin z - \sin y) dz \\ &= -1 - 16 \cos^5 y + 20 \cos^3 y - 5 \cos y - 3 (\sin y) y + 2 \cos^2 y. \end{aligned}$$

Finally, this last expression must be integrated from 0 to π . Thus

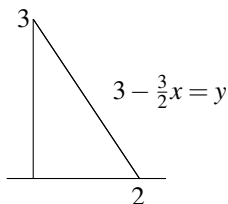
$$\begin{aligned} & \int_0^\pi \int_0^{3y} \int_0^{y+z} \cos(x+y) \, dx \, dz \, dy \\ &= \int_0^\pi \left(-1 - 16\cos^5 y + 20\cos^3 y - 5\cos y - 3(\sin y)y + 2\cos^2 y \right) dy = -3\pi \end{aligned}$$

Example 26.3.5 Here is an iterated integral: $\int_0^2 \int_0^{3-\frac{3}{2}x} \int_0^{x^2} dz \, dy \, dx$. Write as an iterated integral in the order $dz \, dx \, dy$.

The inside integral is just a function of x and y . (In fact, only a function of x .) The order of the last two integrals must be interchanged. Thus the iterated integral which needs to be done in a different order is

$$\int_0^2 \int_0^{3-\frac{3}{2}x} f(x,y) \, dy \, dx.$$

As usual, it is important to draw a picture and then go from there.



Thus this double integral equals

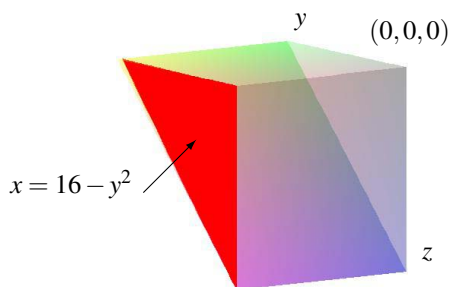
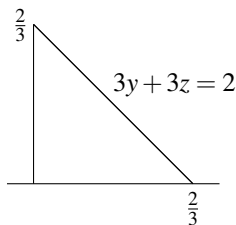
$$\int_0^3 \int_0^{\frac{2}{3}(3-y)} f(x,y) \, dx \, dy.$$

Now substituting in for $f(x,y)$,

$$\int_0^3 \int_0^{\frac{2}{3}(3-y)} \int_0^{x^2} dz \, dx \, dy.$$

Example 26.3.6 Find the volume of the bounded region determined by $3y + 3z = 2$, $x = 16 - y^2$, $y = 0$, $x = 0$.

In the yz plane, the first of the following pictures corresponds to $x = 0$.



Therefore, the outside integrals taken with respect to z and y are of the form

$$\int_0^{\frac{2}{3}} \int_0^{\frac{2}{3}-y} dz dy$$

and now for any choice of (y, z) in the above triangular region, x goes from 0 to $16 - y^2$. Therefore, the iterated integral is

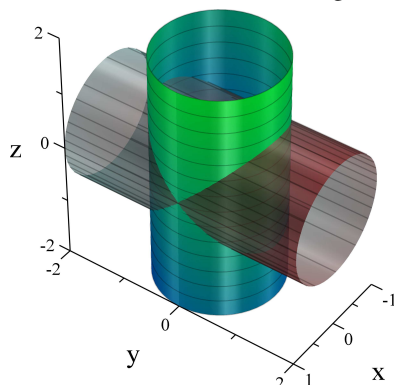
$$\int_0^{\frac{2}{3}} \int_0^{\frac{2}{3}-y} \int_0^{16-y^2} dx dz dy = \frac{860}{243}$$

Example 26.3.7 Find the volume of the region determined by the intersection of the two cylinders, $x^2 + y^2 \leq 1$ and $x^2 + z^2 \leq 1$.

The first listed cylinder intersects the xy plane in the disk, $x^2 + y^2 \leq 1$. What is the volume of the three dimensional region which is between this disk and the two surfaces, $z = \sqrt{1 - x^2}$ and $z = -\sqrt{1 - x^2}$? An iterated integral for the volume is

$$\int_{-1}^1 \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} dz dy dx = \frac{16}{3}.$$

Note that I drew no picture of the three dimensional region. If you are interested, here it is.



One of the cylinders is parallel to the z axis, $x^2 + y^2 \leq 1$ and the other is parallel to the y axis, $x^2 + z^2 \leq 1$. I did not need to be able to draw such a nice picture in order to work this problem. This is the key to doing these. Draw pictures in two dimensions and reason from the two dimensional pictures rather than attempt to wax artistic and consider all three dimensions at once. These problems are hard enough without making them even harder by attempting to be an artist.

26.4 Exercises

1. Find the following iterated integrals.

(a) $\int_{-1}^3 \int_0^{2z} \int_y^{z+1} (x+y) dx dy dz$

(b) $\int_0^1 \int_0^z \int_y^{z^2} (y+z) dx dy dz$

- (c) $\int_0^3 \int_1^x \int_2^{3x-y} \sin(x) \, dz \, dy \, dx$
- (d) $\int_0^1 \int_x^{2x} \int_y^{2y} \, dz \, dy \, dx$
- (e) $\int_2^4 \int_2^{2x} \int_{2y}^x \, dz \, dy \, dx$
- (f) $\int_0^3 \int_0^{2-5x} \int_0^{2-x-2y} 2x \, dz \, dy \, dx$
- (g) $\int_0^2 \int_0^{1-3x} \int_0^{3-3x-2y} x \, dz \, dy \, dx$
- (h) $\int_0^\pi \int_0^{3y} \int_0^{y+z} \cos(x+y) \, dx \, dz \, dy$
- (i) $\int_0^\pi \int_0^{4y} \int_0^{y+z} \sin(x+y) \, dx \, dz \, dy$

2. Fill in the missing limits.

$$\begin{aligned} \int_0^1 \int_0^z \int_0^z f(x, y, z) \, dx \, dy \, dz &= \int_?^? \int_?^? \int_?^? f(x, y, z) \, dx \, dz \, dy, \\ \int_0^1 \int_0^z \int_0^{2z} f(x, y, z) \, dx \, dy \, dz &= \int_?^? \int_?^? \int_?^? f(x, y, z) \, dy \, dz \, dx, \\ \int_0^1 \int_0^z \int_0^z f(x, y, z) \, dx \, dy \, dz &= \int_?^? \int_?^? \int_?^? f(x, y, z) \, dz \, dy \, dx, \\ \int_0^1 \int_{z/2}^z \int_0^{y+z} f(x, y, z) \, dx \, dy \, dz &= \int_?^? \int_?^? \int_?^? f(x, y, z) \, dx \, dz \, dy, \\ \int_4^6 \int_2^6 \int_0^4 f(x, y, z) \, dx \, dy \, dz &= \int_?^? \int_?^? \int_?^? f(x, y, z) \, dz \, dy \, dx. \end{aligned}$$

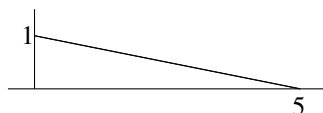
- 3. Find the volume of R where R is the bounded region formed by the plane $\frac{1}{5}x + y + \frac{1}{4}z = 1$ and the planes $x = 0, y = 0, z = 0$.
- 4. Find the volume of R where R is the bounded region formed by the plane $\frac{1}{5}x + \frac{1}{2}y + \frac{1}{4}z = 1$ and the planes $x = 0, y = 0, z = 0$.
- 5. Find the volume of R where R is the bounded region formed by the plane $\frac{1}{5}x + \frac{1}{3}y + \frac{1}{4}z = 1$ and the planes $x = 0, y = 0, z = 0$.
- 6. Find the volume of the bounded region determined by $3y + z = 3, x = 4 - y^2, y = 0, x = 0$.
- 7. Find the volume of the region bounded by $x^2 + y^2 = 16, z = 3x, z = 0$, and $x \geq 0$.
- 8. Find the volume of R where R is the bounded region formed by the plane $\frac{1}{4}x + \frac{1}{2}y + \frac{1}{4}z = 1$ and the planes $x = 0, y = 0, z = 0$.
- 9. Here is an iterated integral: $\int_0^3 \int_0^{3-x} \int_0^{x^2} \, dz \, dy \, dx$. Write as an iterated integral in the following orders: $dz \, dx \, dy, dx \, dz \, dy, dx \, dy \, dz, dy \, dx \, dz, dy \, dz \, dx$.
- 10. Find the volume of the bounded region determined by $2y + z = 3, x = 9 - y^2, y = 0, x = 0, z = 0$.
- 11. Find the volume of the bounded region determined by $y + 2z = 3, x = 9 - y^2, y = 0, x = 0$.
- 12. Find the volume of the bounded region determined by $y + z = 2, x = 3 - y^2, y = 0, x = 0$.
- 13. Find the volume of the region bounded by $x^2 + y^2 = 25, z = x, z = 0$, and $x \geq 0$.
Your answer should be $\frac{250}{3}$.
- 14. Find the volume of the region bounded by $x^2 + y^2 = 9, z = 3x, z = 0$, and $x \geq 0$.

26.4.1 Mass and Density

As an example of the use of triple integrals, consider a solid occupying a set of points $U \subseteq \mathbb{R}^3$ having density ρ . Thus ρ is a function of position and the total mass of the solid equals $\int_U \rho \, dV$. This is just like the two dimensional case. The mass of an infinitesimal chunk of the solid located at \mathbf{x} would be $\rho(\mathbf{x}) \, dV$ and so the total mass is just the sum of all these, $\int_U \rho(\mathbf{x}) \, dV$.

Example 26.4.1 Find the volume of R where R is the bounded region formed by the plane $\frac{1}{5}x + y + \frac{1}{5}z = 1$ and the planes $x = 0, y = 0, z = 0$.

When $z = 0$, the plane becomes $\frac{1}{5}x + y = 1$. Thus the intersection of this plane with the xy plane is this line shown in the following picture.



Therefore, the bounded region is between the triangle formed in the above picture by the x axis, the y axis and the above line and the surface given by $\frac{1}{5}x + y + \frac{1}{5}z = 1$ or $z = 5(1 - (\frac{1}{5}x + y)) = 5 - x - 5y$. Therefore, an iterated integral which yields the volume is

$$\int_0^5 \int_0^{1-\frac{1}{5}x} \int_0^{5-x-5y} dz \, dy \, dx = \frac{25}{6}.$$

Example 26.4.2 Find the mass of the bounded region R formed by the plane $\frac{1}{3}x + \frac{1}{3}y + \frac{1}{5}z = 1$ and the planes $x = 0, y = 0, z = 0$ if the density is $\rho(x, y, z) = z$.

This is done just like the previous example except in this case, there is a function to integrate. Thus the answer is

$$\int_0^3 \int_0^{3-x} \int_0^{5-\frac{5}{3}x-\frac{5}{3}y} z \, dz \, dy \, dx = \frac{75}{8}.$$

Example 26.4.3 Find the total mass of the bounded solid determined by $z = 9 - x^2 - y^2$ and $x, y, z \geq 0$ if the mass is given by $\rho(x, y, z) = z$

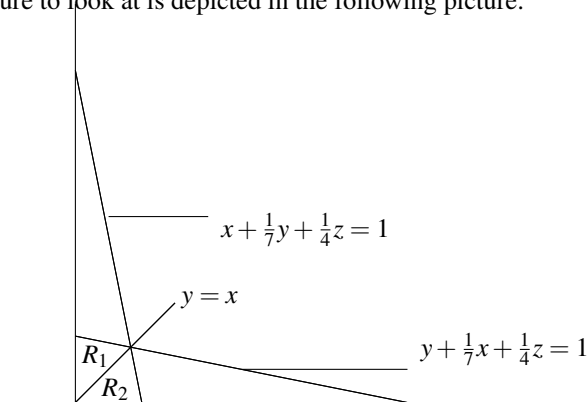
When $z = 0$ the surface $z = 9 - x^2 - y^2$ intersects the xy plane in a circle of radius 3 centered at $(0, 0)$. Since $x, y \geq 0$, it is only a quarter of a circle of interest, the part where both these variables are nonnegative. For each (x, y) inside this quarter circle, z goes from 0 to $9 - x^2 - y^2$. Therefore, the iterated integral is of the form,

$$\int_0^3 \int_0^{\sqrt{9-x^2}} \int_0^{9-x^2-y^2} z \, dz \, dy \, dx = \frac{243}{8}\pi$$

Example 26.4.4 Find the volume of the bounded region determined by $x \geq 0, y \geq 0, z \geq 0$, and $\frac{1}{7}x + y + \frac{1}{4}z = 1$, and $x + \frac{1}{7}y + \frac{1}{4}z = 1$.

When $z = 0$, the plane $\frac{1}{7}x + y + \frac{1}{4}z = 1$ intersects the xy plane in the line whose equation is $\frac{1}{7}x + y = 1$, while the plane, $x + \frac{1}{7}y + \frac{1}{4}z = 1$ intersects the xy plane in the line whose

equation is $x + \frac{1}{7}y = 1$. Furthermore, the two planes intersect when $x = y$ as can be seen from the equations, $x + \frac{1}{7}y = 1 - \frac{z}{4}$ and $\frac{1}{7}x + y = 1 - \frac{z}{4}$ which imply $x = y$. Thus the two dimensional picture to look at is depicted in the following picture.



You see in this picture, the base of the region in the xy plane is the union of the two triangles, R_1 and R_2 . For $(x, y) \in R_1$, z goes from 0 to what it needs to be to be on the plane, $\frac{1}{7}x + y + \frac{1}{4}z = 1$. Thus z goes from 0 to $4(1 - \frac{1}{7}x - y)$. Similarly, on R_2 , z goes from 0 to $4(1 - \frac{1}{7}y - x)$. Therefore, the integral needed is

$$\int_{R_1} \int_0^{4(1-\frac{1}{7}x-y)} dz dV + \int_{R_2} \int_0^{4(1-\frac{1}{7}y-x)} dz dV$$

and now it only remains to consider $\int_{R_1} dV$ and $\int_{R_2} dV$. The point of intersection of these lines shown in the above picture is $(\frac{7}{8}, \frac{7}{8})$ and so an iterated integral is

$$\int_0^{7/8} \int_x^{1-\frac{x}{7}} \int_0^{4(1-\frac{1}{7}x-y)} dz dy dx + \int_0^{7/8} \int_y^{1-\frac{y}{7}} \int_0^{4(1-\frac{1}{7}y-x)} dz dx dy = \frac{7}{6}$$

26.5 Exercises

- Find the volume of the region determined by the intersection of the two cylinders, $x^2 + y^2 \leq 16$ and $y^2 + z^2 \leq 16$.
- Find the volume of the region determined by the intersection of the two cylinders, $x^2 + y^2 \leq 9$ and $y^2 + z^2 \leq 9$.
- Find the volume of the region bounded by $x^2 + y^2 = 4$, $z = 0$, $z = 5 - y$
- Find $\int_0^2 \int_0^{6-2z} \int_{\frac{1}{2}x}^{3-z} (3-z) \cos(y^2) dy dx dz$.
- Find $\int_0^1 \int_0^{18-3z} \int_{\frac{1}{3}x}^{6-z} (6-z) \exp(y^2) dy dx dz$.
- Find $\int_0^2 \int_0^{24-4z} \int_{\frac{1}{4}y}^{6-z} (6-z) \exp(x^2) dx dy dz$.
- Find $\int_0^1 \int_0^{10-2z} \int_{\frac{1}{2}y}^{5-z} \frac{\sin x}{x} dx dy dz$.

Hint: Interchange order of integration.

8. Find the mass of the bounded region R formed by the plane $\frac{1}{4}x + \frac{1}{2}y + \frac{1}{3}z = 1$ and the planes $x = 0, y = 0, z = 0$ if the density is $\rho(x, y, z) = y$
9. Find the mass of the bounded region R formed by the plane $\frac{1}{2}x + \frac{1}{2}y + \frac{1}{4}z = 1$ and the planes $x = 0, y = 0, z = 0$ if the density is $\rho(x, y, z) = z^2$
10. Find the mass of the bounded region R formed by the plane $\frac{1}{4}x + \frac{1}{2}y + \frac{1}{4}z = 1$ and the planes $x = 0, y = 0, z = 0$ if the density is $\rho(x, y, z) = y + z$
11. Find the mass of the bounded region R formed by the plane $\frac{1}{4}x + \frac{1}{2}y + \frac{1}{5}z = 1$ and the planes $x = 0, y = 0, z = 0$ if the density is $\rho(x, y, z) = y$
12. Find $\int_0^1 \int_0^{12-4z} \int_{\frac{1}{4}y}^{3-z} \frac{\sin x}{x} dx dy dz$.
13. Find $\int_0^{20} \int_0^2 \int_{\frac{1}{3}y}^{6-z} \frac{\sin x}{x} dx dz dy + \int_{20}^{30} \int_0^{6-\frac{1}{5}y} \int_{\frac{1}{5}y}^{6-z} \frac{\sin x}{x} dx dz dy$.
14. Find the volume of the bounded region determined by $x \geq 0, y \geq 0, z \geq 0$, and $\frac{1}{2}x + y + \frac{1}{2}z = 1$, and $x + \frac{1}{2}y + \frac{1}{2}z = 1$.
15. Find the volume of the bounded region determined by $x \geq 0, y \geq 0, z \geq 0$, and $\frac{1}{7}x + y + \frac{1}{3}z = 1$, and $x + \frac{1}{7}y + \frac{1}{3}z = 1$.
16. Find an iterated integral for the volume of the region between the graphs of $z = x^2 + y^2$ and $z = 2(x + y)$.
17. Find the volume of the region which lies between $z = x^2 + y^2$ and the plane $z = 4$.
18. The base of a solid is the region in the xy plane between the curves $y = x^2$ and $y = 1$. The top of the solid is the plane $z = 2 - x$. Find the volume of the solid.
19. The base of a solid is in the xy plane and is bounded by the lines $y = x, y = 1 - x$, and $y = 0$. The top of the solid is $z = 3 - y$. Find its volume.
20. The base of a solid is in the xy plane and is bounded by the lines $x = 0, x = \pi, y = 0$, and $y = \sin x$. The top of this solid is $z = x$. Find the volume of this solid.

Chapter 27

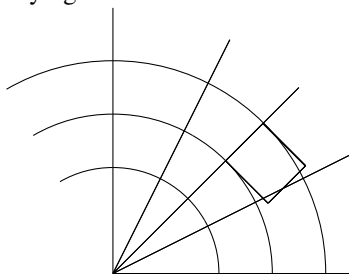
The Integral in Other Coordinates

27.1 Polar Coordinates

Recall the relation between the rectangular coordinates and polar coordinates is

$$\mathbf{x}(r, \theta) \equiv \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} r \cos(\theta) \\ r \sin(\theta) \end{pmatrix}, \quad r \geq 0, \quad \theta \in [0, 2\pi)$$

Now consider the part of grid obtained by fixing θ at various values and varying r and then by fixing r at various values and varying θ .



The idea is that these lines obtained by fixing one or the other coordinate are very close together, much closer than drawn and so we would expect the area of one of the little curvy quadrilaterals to be close to the area of the parallelogram shown. Consider this parallelogram. The two sides originating at the intersection of two of the grid lines as shown are approximately equal to

$$\mathbf{x}_r(r, \theta) dr, \quad \mathbf{x}_\theta(r, \theta) d\theta$$

where dr and $d\theta$ are the respective small changes in the variables r and θ . Thus the area of one of those little curvy shapes should be approximately equal to

$$|\mathbf{x}_r(r, \theta) dr \times \mathbf{x}_\theta(r, \theta) d\theta|$$

by the geometric description of the cross product. These vectors are extended as 0 in the

third component in order to take the cross product. This reduces to

$$dA = \left| \det \begin{pmatrix} \cos(\theta) & -r \sin(\theta) \\ \sin(\theta) & r \cos(\theta) \end{pmatrix} \right| dr d\theta = r dr d\theta$$

which is the increment of area in polar coordinates, taking the place of $dx dy$. The integral is really about taking the value of the function integrated multiplied by dA and adding these products. Here is an example.

Example 27.1.1 Find the area of a circle of radius a .

The variable r goes from 0 to a and the angle θ goes from 0 to 2π . Therefore, the area is

$$\int_D dA = \int_0^{2\pi} \int_0^a r dr d\theta = \pi a^2$$

Example 27.1.2 The density equals r . Find the total mass of a disk of radius a .

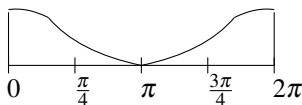
This is easy to do in polar coordinates. The disk involved has θ going from 0 to 2π and r from 0 to a . Therefore, the integral to work is just

$$\int_0^{2\pi} \int_0^a \overbrace{r dr d\theta}^{dA} = \frac{2}{3} \pi a^3$$

Notice how in these examples the circular disk is really a rectangle $[0, 2\pi] \times [0, a]$. This is why polar coordinates are so useful. The next example was worked earlier from a different point of view.

Example 27.1.3 Find the area of the inside of the cardioid $r = 1 + \cos \theta$, $\theta \in [0, 2\pi]$.

How would you go about setting this up in rectangular coordinates? It would be very hard if not impossible, but is easy in polar coordinates. This is because in polar coordinates the region integrated over is the region below the curve in the following picture. It is one of those regions which is simple to integrate over. The graph of the top is $r = 1 + \cos \theta$. However, the graph of the cardioid in rectangular coordinates is not at all simple. See the material on polar coordinates where graphs of cardioids were provided in rectangular coordinates.



The integral is

$$\int_0^{2\pi} \int_0^{1+\cos(\theta)} r dr d\theta = \frac{3}{2} \pi$$

Example 27.1.4 Let R denote the inside of the cardioid $r = 1 + \cos \theta$ for $\theta \in [0, 2\pi]$. Find

$$\int_R x dA$$

Here the convenient increment of area is $rdrd\theta$ and so the integral is

$$\int_0^{2\pi} \int_0^{1+\cos(\theta)} x r dr d\theta$$

Now you need to change x to the right coordinates. Thus the integral equals

$$\int_0^{2\pi} \int_0^{1+\cos(\theta)} (r \cos(\theta)) r dr d\theta = \frac{5}{4}\pi$$

A case where this sort of problem occurs is when you find the mass of a plate given the density.

Definition 27.1.5 Suppose a material occupies a region of the plane R . The density λ is a nonnegative function of position with the property that if $B \subseteq R$, then the mass of B is given by $\int_B \lambda dA$. In particular, this is true of $B = R$.

Example 27.1.6 Let R denote the inside of the polar curve $r = 2 + \sin \theta$. Let $\lambda = 3 + x$. Find the total mass of R .

As above, this is

$$\int_0^{2\pi} \int_0^{2+\sin(\theta)} (3 + r \cos(\theta)) r dr d\theta = \frac{27}{2}\pi$$

27.2 Exercises

1. Sketch a graph in polar coordinates of $r = 2 + \sin(\theta)$ and find the area of the enclosed region.
2. Sketch a graph in polar coordinates of $r = \sin(4\theta)$ and find the area of the region enclosed. **Hint:** In this case, you need to worry and fuss about $r < 0$.
3. Suppose the density is $\lambda(x, y) = 2 - x$ and the region is the interior of the cardioid $r = 1 + \cos \theta$. Find the total mass.
4. Suppose the density is $\lambda = 4 - x - y$ and find the mass of the plate which is between the concentric circles $r = 1$ and $r = 2$.
5. Suppose the density is $\lambda = 4 - x - y$ and find the mass of the plate which is inside the polar graph of $r = 1 + \sin(\theta)$.
6. Suppose the density is $2 + x$. Find the mass of the plate which is the inside of the polar curve $r = \sin(2\theta)$. **Hint:** This is one of those fussy things with negative radius.
7. The area density of a plate is given by $\lambda = 1 + x$ and the plate occupies the inside of the cardioid $r = 1 + \cos \theta$. Find its mass.
8. The moment about the x axis of a plate with density λ occupying the region R is defined as $m_y = \int_R y \lambda dA$. The moment about the y axis of the same plate is $m_x = \int_R x \lambda dA$. If $\lambda = 2 - x$, find the moments about the x and y axes of the plate inside $r = 2 + \sin(\theta)$.

9. Using the above problem, find the moments about the x and y axes of a plate having density $1 + x$ for the plate which is the inside of the cardioid $r = 1 + \cos \theta$.
10. Use the same plate as the above but this time, let the density be $(2 + x + y)$. Find the moments.
11. Let $D = \{(x, y) : x^2 + y^2 \leq 25\}$. Find $\int_D e^{25x^2 + 25y^2} dx dy$. **Hint:** This is an integral of the form $\int_D f(x, y) dA$. Write in polar coordinates and it will be fairly easy.
12. Let $D = \{(x, y) : x^2 + y^2 \leq 16\}$. Find $\int_D \cos(9x^2 + 9y^2) dx dy$. **Hint:** This is an integral of the form $\int_D f(x, y) dA$. Write in polar coordinates and it will be fairly easy.
13. Derive a formula for area between two polar graphs using the increment of area of polar coordinates.
14. Use polar coordinates to evaluate the following integral. Here S is given in terms of the polar coordinates. $\int_S \sin(2x^2 + 2y^2) dV$ where $r \leq 2$ and $0 \leq \theta \leq \frac{3}{2}\pi$.
15. Find $\int_S e^{2x^2 + 2y^2} dV$ where S is given in terms of the polar coordinates $r \leq 2$ and $0 \leq \theta \leq \pi$.
16. Find $\int_S \frac{y}{x} dV$ where S is described in polar coordinates as $1 \leq r \leq 2$ and $0 \leq \theta \leq \pi/4$.
17. Find $\int_S \left(\left(\frac{y}{x} \right)^2 + 1 \right) dV$ where S is given in polar coordinates as $1 \leq r \leq 2$ and $0 \leq \theta \leq \frac{1}{6}\pi$.
18. A right circular cone has a base of radius 2 and a height equal to 2. Use polar coordinates to find its volume.
19. Now suppose in the above problem, it is not really a cone but instead $z = 2 - \frac{1}{2}r^2$. Find its volume.

27.3 Cylindrical and Spherical Coordinates

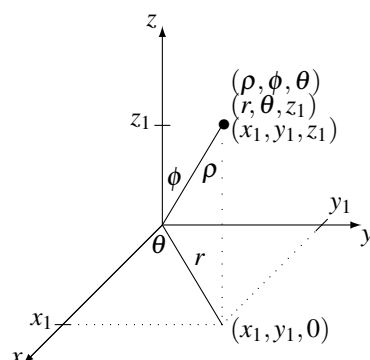
Cylindrical coordinates are defined as follows.

$$\begin{aligned} \mathbf{x}(r, \theta, z) &\equiv \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} r \cos(\theta) \\ r \sin(\theta) \\ z \end{pmatrix}, \\ r &\geq 0, \theta \in [0, 2\pi), z \in \mathbb{R} \end{aligned}$$

Spherical coordinates are a little harder. These are given by

$$\begin{aligned} \mathbf{x}(\rho, \theta, \phi) &\equiv \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \rho \sin(\phi) \cos(\theta) \\ \rho \sin(\phi) \sin(\theta) \\ \rho \cos(\phi) \end{pmatrix}, \\ \rho &\geq 0, \theta \in [0, 2\pi), \phi \in [0, \pi] \end{aligned}$$

The following picture relates the various coordinates.

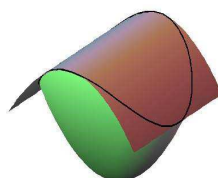


In this picture, ρ is the distance between the origin, the point whose Cartesian coordinates are $(0,0,0)$ and the point indicated by a dot and labelled as (x_1, y_1, z_1) , (r, θ, z_1) , and (ρ, ϕ, θ) . The angle between the positive z axis and the line between the origin and the point indicated by a dot is denoted by ϕ , and θ is the angle between the positive x axis and the line joining the origin to the point $(x_1, y_1, 0)$ as shown, while r is the length of this line. Thus $r = \rho \sin(\phi)$ and is the usual polar coordinate while θ is the other polar coordinate. Letting z_1 denote the usual z coordinate of a point in three dimensions, like the one shown as a dot, (r, θ, z_1) are the cylindrical coordinates of the dotted point. The spherical coordinates are determined by (ρ, ϕ, θ) . When ρ is specified, this indicates that the point of interest is on some sphere of radius ρ which is centered at the origin. Then when ϕ is given, the location of the point is narrowed down to a circle of “latitude” and finally, θ determines which point is on this circle by specifying a circle of “longitude”. Let $\phi \in [0, \pi]$, $\theta \in [0, 2\pi)$, and $\rho \in [0, \infty)$. The picture shows how to relate these new coordinate systems to Cartesian coordinates. Note that θ is the same in the two coordinate systems and that $\rho \sin \phi = r$.

27.3.1 Volume and Integrals in Cylindrical Coordinates

The increment of three dimensional volume in cylindrical coordinates is $dV = r dr d\theta dz$. It is just a chunk of two dimensional area $r dr d\theta$ times the height dz which gives three dimensional volume. Here is an example.

Example 27.3.1 Find the volume of the three dimensional region between the graphs of $z = 4 - 2y^2$ and $z = 4x^2 + 2y^2$.



Where do the two surfaces intersect? This happens when $4x^2 + 2y^2 = 4 - 2y^2$ which is the curve in the xy plane, $x^2 + y^2 = 1$. Thus (x, y) is on the inside of this circle while z goes from $4x^2 + 2y^2$ to $4 - 2y^2$. Denoting the unit disk by D , the desired integral is

$$\int_D \int_{4x^2+2y^2}^{4-2y^2} dz dA$$

I will use the dA which corresponds to polar coordinates so this will then be in cylindrical coordinates. Thus the above equals

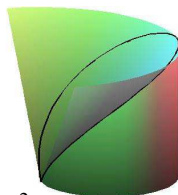
$$\int_0^{2\pi} \int_0^1 \int_{4(r^2 \cos^2(\theta)) + 2(r^2 \sin^2(\theta))}^{4-2(r^2 \sin^2(\theta))} dz r dr d\theta = 2\pi$$

Note this is really not much different than simply using polar coordinates to integrate the difference of the two values of z . This is

$$\begin{aligned} \int_D 4 - 2y^2 - (4x^2 + 2y^2) dA &= \int_D (4 - 4r^2) dA \\ &= \int_0^{2\pi} \int_0^1 (4 - 4r^2) r dr d\theta = 2\pi \end{aligned}$$

Here is another example.

Example 27.3.2 Find the volume of the three dimensional region between the graphs of $z = 0$, $z = \sqrt{x^2 + y^2}$, and the cylinder $(x - 1)^2 + y^2 = 1$.

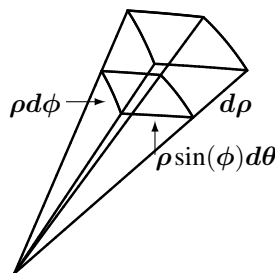
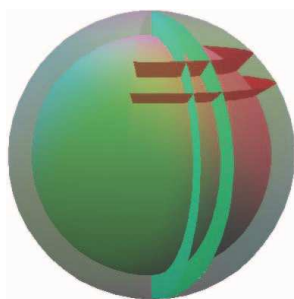


Consider the cylinder. It reduces to $r^2 = 2r \cos \theta$ or more simply $r = 2 \cos \theta$. This is the graph of a circle having radius 1 and centered at $(1, 0)$. Therefore, $\theta \in [-\pi/2, \pi/2]$. It follows that the cylindrical coordinate description of this volume is

$$\int_{-\pi/2}^{\pi/2} \int_0^{2 \cos \theta} \int_0^r dz r dr d\theta = \frac{32}{9}$$

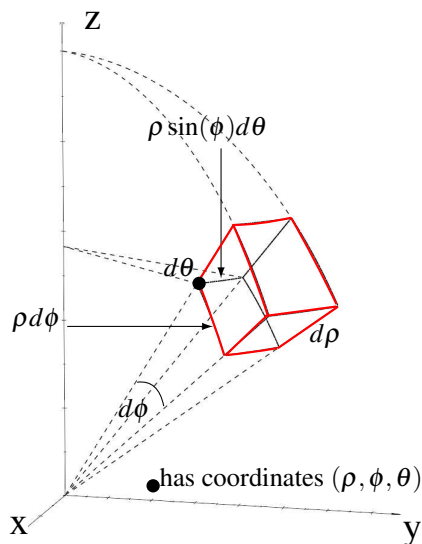
27.3.2 Volume and Integrals in Spherical Coordinates

What is the increment of volume in spherical coordinates? There are two ways to see what this is, through art and through a systematic procedure. First consider art. Here is a picture.



In the picture there are two concentric spheres formed by making ρ two different constants and surfaces which correspond to θ assuming two different constants and ϕ assuming

two different constants. These intersecting surfaces form the little box in the picture. Here is a more detailed blow up of the little box.



What is the volume of this little box? Length $\approx \rho d\phi$, width $\approx \rho \sin(\phi) d\theta$, height $\approx d\rho$ and so the volume increment for spherical coordinates is

$$dV = \rho^2 \sin(\phi) d\rho d\theta d\phi$$

Now what is really going on? Consider the dot in the picture of the little box. Fixing θ and ϕ at their values at this point and differentiating with respect to ρ leads to a little vector of the form

$$\begin{pmatrix} \sin(\phi) \cos(\theta) \\ \sin(\phi) \sin(\theta) \\ \cos(\phi) \end{pmatrix} d\rho$$

which points out from the surface of the sphere. Next keeping ρ and θ constant and differentiating only with respect to ϕ leads to an infinitesimal vector in the direction of a line of longitude,

$$\begin{pmatrix} \rho \cos(\phi) \cos(\theta) \\ \rho \cos(\phi) \sin(\theta) \\ -\rho \sin(\phi) \end{pmatrix} d\phi$$

and finally keeping ρ and ϕ constant and differentiating with respect to θ leads to the third infinitesimal vector which points in the direction of a line of latitude.

$$\begin{pmatrix} -\rho \sin(\phi) \sin(\theta) \\ \rho \sin(\phi) \cos(\theta) \\ 0 \end{pmatrix} d\theta$$

To find the increment of volume, we just need to take the absolute value of the determinant which has these vectors as columns, (Remember this is the absolute value of the box product.) exactly as was the case for polar coordinates. This will also yield

$$dV = \rho^2 \sin(\phi) d\rho d\theta d\phi.$$

However, in contrast to the drawing of pictures, this procedure is completely general and will handle all curvilinear coordinate systems and in any dimension. This is discussed more later.

Example 27.3.3 Find the volume of a ball, B_R of radius R . Then find $\int_{B_R} z^2 dV$ where z is the rectangular z coordinate of a point.

In this case, $U = (0, R] \times [0, \pi] \times [0, 2\pi)$ and use spherical coordinates. Then this yields a set in \mathbb{R}^3 which clearly differs from the ball of radius R only by a set having volume equal to zero. It leaves out the point at the origin is all. Therefore, the volume of the ball is

$$\begin{aligned} \int_{B_R} 1 dV &= \int_U \rho^2 \sin \phi dV \\ &= \int_0^R \int_0^\pi \int_0^{2\pi} \rho^2 \sin \phi d\theta d\phi d\rho = \frac{4}{3} R^3 \pi. \end{aligned}$$

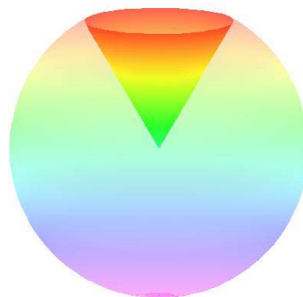
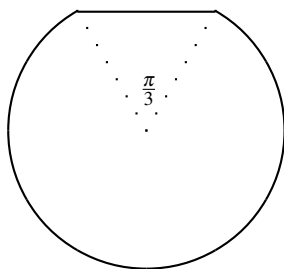
The reason this was effortless, is that the ball, B_R is realized as a box in terms of the spherical coordinates. Remember what was pointed out earlier about setting up iterated integrals over boxes.

As for the integral, it is no harder to set up. You know from the transformation equations that $z = \rho \cos \phi$. Then you want

$$\int_{B_R} z dV = \int_0^R \int_0^\pi \int_0^{2\pi} (\rho \cos(\phi))^2 \rho^2 \sin \phi d\theta d\phi d\rho = \frac{4}{15} \pi R^5$$

This will be pretty easy also although somewhat more messy because the function you are integrating is not just 1 as it is when you find the volume.

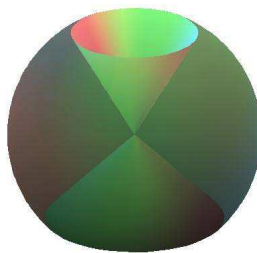
Example 27.3.4 A cone is cut out of a ball of radius R as shown in the following picture, the diagram on the left being a side view. The angle of the cone is $\pi/3$. Find the volume of what is left.



Use spherical coordinates. This volume is then

$$\int_{\pi/6}^\pi \int_0^{2\pi} \int_0^R \rho^2 \sin(\phi) d\rho d\theta d\phi = \frac{2}{3} \pi R^3 + \frac{1}{3} \sqrt{3} \pi R^3$$

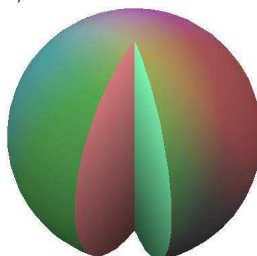
Now change the example a little by cutting out a cone at the bottom which has an angle of $\pi/2$ as shown. What is the volume of what is left?



This time you would have the volume equals

$$\int_{\pi/6}^{3\pi/4} \int_0^{2\pi} \int_0^R \rho^2 \sin(\phi) d\rho d\theta d\phi = \frac{1}{3}\sqrt{2}\pi R^3 + \frac{1}{3}\sqrt{3}\pi R^3$$

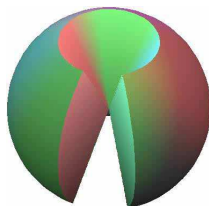
Example 27.3.5 Next suppose the ball of radius R is a sort of an orange and you remove a slice as shown in the picture. What is the volume of what is left? Assume the slice is formed by the two half planes $\theta = 0$ and $\theta = \pi/4$.



Using spherical coordinates, this gives for the volume

$$\int_0^\pi \int_{\pi/4}^{2\pi} \int_0^R \rho^2 \sin(\phi) d\rho d\theta d\phi = \frac{7}{6}\pi R^3$$

Example 27.3.6 Now remove the same two cones as in the above examples along with the same slice and find the volume of what is left. Next, if R is the region just described, find $\int_R x dV$.



This time you need

$$\int_{\pi/6}^{3\pi/4} \int_{\pi/4}^{2\pi} \int_0^R \rho^2 \sin(\phi) d\rho d\theta d\phi = \frac{7}{24}\sqrt{2}\pi R^3 + \frac{7}{24}\sqrt{3}\pi R^3$$

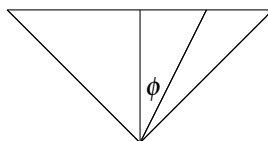
As to the integral, it equals

$$\int_{\pi/6}^{3\pi/4} \int_{\pi/4}^{2\pi} \int_0^R (\rho \sin(\phi) \cos(\theta)) \rho^2 \sin(\phi) d\rho d\theta d\phi = -\frac{1}{192}\sqrt{2}R^4 (7\pi + 3\sqrt{3} + 6)$$

This is because, in terms of spherical coordinates, $x = \rho \sin(\phi) \cos(\theta)$.

Example 27.3.7 Set up the integrals to find the volume of the cone $0 \leq z \leq 4, z = \sqrt{x^2 + y^2}$. Next, if R is the region just described, find $\int_R z dV$.

This is entirely the wrong coordinate system to use for this problem but it is a good exercise. Here is a side view.



You need to figure out what ρ is as a function of ϕ which goes from 0 to $\pi/4$. You should get

$$\int_0^{2\pi} \int_0^{\pi/4} \int_0^{4 \sec(\phi)} \rho^2 \sin(\phi) d\rho d\phi d\theta = \frac{64}{3}\pi$$

As to $\int_R z dV$, it equals

$$\int_0^{2\pi} \int_0^{\pi/4} \int_0^{4 \sec(\phi)} \overbrace{\rho \cos(\phi)}^z \rho^2 \sin(\phi) d\rho d\phi d\theta = 64\pi$$

Example 27.3.8 Find the volume element for cylindrical coordinates.

In cylindrical coordinates,

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} r \cos \theta \\ r \sin \theta \\ z \end{pmatrix}$$

Therefore, the Jacobian determinant is

$$\det \begin{pmatrix} \cos \theta & -r \sin \theta & 0 \\ \sin \theta & r \cos \theta & 0 \\ 0 & 0 & 1 \end{pmatrix} = r.$$

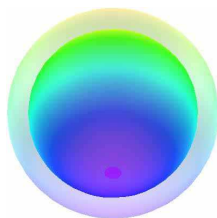
It follows the volume element in cylindrical coordinates is $r d\theta dr dz$.

Example 27.3.9 In the cone of Example 27.3.7 set up the integrals for finding the volume in cylindrical coordinates.

This is a better coordinate system for this example than spherical coordinates. This time you should get

$$\int_0^{2\pi} \int_0^4 \int_r^4 r dz dr d\theta = \frac{64}{3}\pi$$

Example 27.3.10 This example uses spherical coordinates to verify an important conclusion about gravitational force. Let the hollow sphere, H be defined by $a^2 < x^2 + y^2 + z^2 < b^2$



and suppose this hollow sphere has constant density taken to equal 1. Now place a unit mass at the point $(0, 0, z_0)$ where $|z_0| \in [a, b]$. Show that the force of gravity acting on this unit mass is $\left(\alpha G \int_H \frac{(z-z_0)}{[x^2+y^2+(z-z_0)^2]^{3/2}} dV \right) \mathbf{k}$ and then show that if $|z_0| > b$ then the force of gravity acting on this point mass is the same as if the entire mass of the hollow sphere were placed at the origin, while if $|z_0| < a$, the total force acting on the point mass from gravity equals zero. Here G is the gravitation constant and α is the density. In particular, this shows that the force a planet exerts on an object is as though the entire mass of the planet were situated at its center¹.

Without loss of generality, assume $z_0 > 0$. Let dV be a little chunk of material located at the point (x, y, z) of H the hollow sphere. Then according to Newton's law of gravity, the force this small chunk of material exerts on the given point mass equals

$$\frac{x\mathbf{i} + y\mathbf{j} + (z - z_0)\mathbf{k}}{|x\mathbf{i} + y\mathbf{j} + (z - z_0)\mathbf{k}|} \frac{1}{(x^2 + y^2 + (z - z_0)^2)} G\alpha dV =$$

$$(x\mathbf{i} + y\mathbf{j} + (z - z_0)\mathbf{k}) \frac{1}{(x^2 + y^2 + (z - z_0)^2)^{3/2}} G\alpha dV$$

Therefore, the total force is

$$\int_H (x\mathbf{i} + y\mathbf{j} + (z - z_0)\mathbf{k}) \frac{1}{(x^2 + y^2 + (z - z_0)^2)^{3/2}} G\alpha dV.$$

By the symmetry of the sphere, the \mathbf{i} and \mathbf{j} components will cancel out when the integral is taken. This is because there is the same amount of stuff for negative x and y as there is for positive x and y . Hence what remains is

$$\alpha G \mathbf{k} \int_H \frac{(z - z_0)}{[x^2 + y^2 + (z - z_0)^2]^{3/2}} dV$$

as claimed. Now for the interesting part, the integral is evaluated. In spherical coordinates this integral is.

$$\int_0^{2\pi} \int_a^b \int_0^\pi \frac{(\rho \cos \phi - z_0) \rho^2 \sin \phi}{(\rho^2 + z_0^2 - 2\rho z_0 \cos \phi)^{3/2}} d\phi d\rho d\theta. \quad (27.1)$$

Rewrite the inside integral and use integration by parts to obtain this inside integral equals

$$\frac{1}{2z_0} \int_0^\pi (\rho^2 \cos \phi - \rho z_0) \frac{(2z_0 \rho \sin \phi)}{(\rho^2 + z_0^2 - 2\rho z_0 \cos \phi)^{3/2}} d\phi =$$

$$\frac{1}{2z_0} \left(-2 \frac{-\rho^2 - \rho z_0}{\sqrt{(\rho^2 + z_0^2 + 2\rho z_0)}} + 2 \frac{\rho^2 - \rho z_0}{\sqrt{(\rho^2 + z_0^2 - 2\rho z_0)}} \right)$$

¹This was shown by Newton in 1685 and allowed him to assert his law of gravitation applied to the planets as though they were point masses. It was a major accomplishment.

$$- \int_0^\pi 2\rho^2 \frac{\sin \phi}{\sqrt{(\rho^2 + z_0^2 - 2\rho z_0 \cos \phi)}} d\phi \Bigg). \quad (27.2)$$

There are some cases to consider here.

First suppose $z_0 < a$ so the point is on the inside of the hollow sphere and it is always the case that $\rho > z_0$. Then in this case, the two first terms reduce to

$$\frac{2\rho(\rho + z_0)}{\sqrt{(\rho + z_0)^2}} + \frac{2\rho(\rho - z_0)}{\sqrt{(\rho - z_0)^2}} = \frac{2\rho(\rho + z_0)}{(\rho + z_0)} + \frac{2\rho(\rho - z_0)}{\rho - z_0} = 4\rho$$

and so the expression in 27.2 equals

$$\begin{aligned} & \frac{1}{2z_0} \left(4\rho - \int_0^\pi 2\rho^2 \frac{\sin \phi}{\sqrt{(\rho^2 + z_0^2 - 2\rho z_0 \cos \phi)}} d\phi \right) \\ &= \frac{1}{2z_0} \left(4\rho - \frac{1}{z_0} \int_0^\pi \rho \frac{2\rho z_0 \sin \phi}{\sqrt{(\rho^2 + z_0^2 - 2\rho z_0 \cos \phi)}} d\phi \right) \\ &= \frac{1}{2z_0} \left(4\rho - \frac{2\rho}{z_0} (\rho^2 + z_0^2 - 2\rho z_0 \cos \phi)^{1/2} \Big|_0^\pi \right) \\ &= \frac{1}{2z_0} \left(4\rho - \frac{2\rho}{z_0} [(\rho + z_0) - (\rho - z_0)] \right) = 0. \end{aligned}$$

Therefore, in this case the inner integral of 27.1 equals zero and so the original integral will also be zero.

The other case is when $z_0 > b$ and so it is always the case that $z_0 > \rho$. In this case the first two terms of 27.2 are

$$\frac{2\rho(\rho + z_0)}{\sqrt{(\rho + z_0)^2}} + \frac{2\rho(\rho - z_0)}{\sqrt{(\rho - z_0)^2}} = \frac{2\rho(\rho + z_0)}{(\rho + z_0)} + \frac{2\rho(\rho - z_0)}{z_0 - \rho} = 0.$$

Therefore in this case, 27.2 equals

$$\begin{aligned} & \frac{1}{2z_0} \left(- \int_0^\pi 2\rho^2 \frac{\sin \phi}{\sqrt{(\rho^2 + z_0^2 - 2\rho z_0 \cos \phi)}} d\phi \right) \\ &= \frac{-\rho}{2z_0^2} \left(\int_0^\pi \frac{2\rho z_0 \sin \phi}{\sqrt{(\rho^2 + z_0^2 - 2\rho z_0 \cos \phi)}} d\phi \right) \end{aligned}$$

which equals

$$\frac{-\rho}{z_0^2} \left((\rho^2 + z_0^2 - 2\rho z_0 \cos \phi)^{1/2} \Big|_0^\pi \right) = \frac{-\rho}{z_0^2} [(\rho + z_0) - (z_0 - \rho)] = -\frac{2\rho^2}{z_0^2}.$$

Thus the inner integral of 27.1 reduces to the above simple expression. Therefore, 27.1 equals

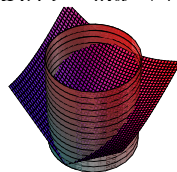
$$\int_0^{2\pi} \int_a^b \left(-\frac{2}{z_0^2} \rho^2 \right) d\rho d\theta = -\frac{4}{3} \pi \frac{b^3 - a^3}{z_0^2}$$

and so

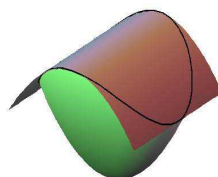
$$\begin{aligned} & \alpha G k \int_H \frac{(z - z_0)}{[x^2 + y^2 + (z - z_0)^2]^{3/2}} dV \\ &= \alpha G k \left(-\frac{4}{3} \pi \frac{b^3 - a^3}{z_0^2} \right) = -kG \frac{\text{total mass}}{z_0^2}. \end{aligned}$$

27.4 Exercises

- Find the volume of the region bounded by $z = 0$, $x^2 + (y - 2)^2 = 4$, and $z = \sqrt{x^2 + y^2}$.



- Find the volume of the region $z \geq 0$, $x^2 + y^2 \leq 4$, and $z \leq 4 - \sqrt{x^2 + y^2}$.
- Find the volume of the region which is between the surfaces $z = 5y^2 + 9x^2$ and $z = 9 - 4y^2$.



- Find the volume of the region which is between $z = x^2 + y^2$ and $z = 5 - 4x$. **Hint:** You might want to change variables at some point.
- The ice cream in a sugar cone is described in spherical coordinates by $\rho \in [0, 10]$, $\phi \in [0, \frac{1}{3}\pi]$, $\theta \in [0, 2\pi]$. If the units are in centimeters, find the total volume in cubic centimeters of this ice cream.
- Find the volume between $z = 3 - x^2 - y^2$ and $z = 2\sqrt{x^2 + y^2}$.
- A ball of radius 3 is placed in a drill press and a hole of radius 2 is drilled out with the center of the hole a diameter of the ball. What is the volume of the material which remains?
- Find the volume of the cone defined by $z \in [0, 4]$ having angle $\pi/2$. Use spherical coordinates.
- A ball of radius 9 has density equal to $\sqrt{x^2 + y^2 + z^2}$ in rectangular coordinates. The top of this ball is sliced off by a plane of the form $z = 2$. Write integrals for the mass of what is left. In spherical coordinates and in cylindrical coordinates.

10. A ball of radius 4 has a cone taken out of the top which has an angle of $\pi/2$ and then a cone taken out of the bottom which has an angle of $\pi/3$. Then a slice, $\theta \in [0, \pi/4]$ is removed. What is the volume of what is left?
11. In Example 27.3.10 on Page 564 check out all the details by working the integrals to be sure the steps are right.
12. What if the hollow sphere in Example 27.3.10 were in two dimensions and everything, including Newton's law still held? Would similar conclusions hold? Explain.
13. Convert the following integrals into integrals involving cylindrical coordinates and then evaluate them.

$$(a) \int_{-2}^2 \int_0^{\sqrt{4-x^2}} \int_0^x xy dz dy dx$$

$$(b) \int_{-1}^1 \int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} \int_0^{x+y} dz dx dy$$

$$(c) \int_0^1 \int_0^{\sqrt{1-x^2}} \int_x^1 dz dy dx$$

$$(d) \text{ For } a > 0, \int_{-a}^a \int_{-\sqrt{a^2-x^2}}^{\sqrt{a^2-x^2}} \int_{-\sqrt{a^2-x^2-y^2}}^{\sqrt{a^2-x^2-y^2}} dz dy dx$$

$$(e) \int_{-1}^1 \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \int_{-\sqrt{4-x^2-y^2}}^{\sqrt{4-x^2-y^2}} dz dy dx$$

14. Convert the following integrals into integrals involving spherical coordinates and then evaluate them.

$$(a) \int_{-a}^a \int_{-\sqrt{a^2-x^2}}^{\sqrt{a^2-x^2}} \int_{-\sqrt{a^2-x^2-y^2}}^{\sqrt{a^2-x^2-y^2}} dz dy dx$$

$$(b) \int_{-1}^1 \int_0^{\sqrt{1-x^2}} \int_{-\sqrt{1-x^2-y^2}}^{\sqrt{1-x^2-y^2}} dz dy dx$$

$$(c) \int_{-\sqrt{2}}^{\sqrt{2}} \int_{-\sqrt{2-x^2}}^{\sqrt{2-x^2}} \int_{\sqrt{x^2+y^2}}^{\sqrt{4-x^2-y^2}} dz dy dx$$

$$(d) \int_{-\sqrt{3}}^{\sqrt{3}} \int_{-\sqrt{3-x^2}}^{\sqrt{3-x^2}} \int_1^{\sqrt{4-x^2-y^2}} dz dy dx$$

$$(e) \int_{-1}^1 \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \int_{-\sqrt{4-x^2-y^2}}^{\sqrt{4-x^2-y^2}} dz dy dx$$

27.5 The General Procedure

As mentioned above, the fundamental concept of an integral is a sum of things of the form $f(\mathbf{x}) dV$ where dV is an “infinitesimal” chunk of volume located at the point \mathbf{x} . Up to now, this infinitesimal chunk of volume has had the form of a box with sides dx_1, \dots, dx_p so $dV = dx_1 dx_2 \cdots dx_p$ but its form is not important. It could just as well be an infinitesimal parallelepiped for example. In what follows, this is what it will be.

First recall the definition of a parallelepiped.

Definition 27.5.1 Let $\mathbf{u}_1, \dots, \mathbf{u}_p$ be vectors in \mathbb{R}^k . The parallelepiped determined by these vectors will be denoted by $P(\mathbf{u}_1, \dots, \mathbf{u}_p)$ and it is defined as

$$P(\mathbf{u}_1, \dots, \mathbf{u}_p) \equiv \left\{ \sum_{j=1}^p s_j \mathbf{u}_j : s_j \in [0, 1] \right\}.$$

Now define the volume of this parallelepiped.

$$\text{volume of } P(\mathbf{u}_1, \dots, \mathbf{u}_p) \equiv (\det(\mathbf{u}_i \cdot \mathbf{u}_j))^{1/2}.$$

To justify this definition, recall that if each vector is in \mathbb{R}^p , the volume of this parallelepiped is $|\det(\begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_p \end{pmatrix})|$. In three dimensions it is $|\mathbf{u}_1 \times \mathbf{u}_2 \cdot \mathbf{u}_3|$. Thus, making the obvious generalization and using $\det(A) = \det(A^T)$, the volume in p dimensions is

$$\begin{aligned} & \left(\det \left(\begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_p \end{pmatrix}^T \det \begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_p \end{pmatrix} \right) \right)^{1/2} \\ &= \det \left[\left(\begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_p \end{pmatrix}^T \begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_p \end{pmatrix} \right) \right]^{1/2} \end{aligned}$$

The ij^{th} entry of the matrix on the inside of $[\cdot]$ is $\mathbf{u}_i \cdot \mathbf{u}_j$ and this is why this definition corresponds to earlier material. Definition 27.5.1 continues to hold in more general settings including the case where the vectors are in \mathbb{R}^q and you have a p dimensional parallelepiped. See Section 20.3.

The dot product is used to determine this volume of a parallelepiped spanned by the given vectors and you should note that it is only the dot product that matters. Let

$$x = f_1(u_1, u_2, u_3), y = f_2(u_1, u_2, u_3), z = f_3(u_1, u_2, u_3) \quad (27.3)$$

where $\mathbf{u} \in U$ an open set in \mathbb{R}^3 and corresponding to such a $\mathbf{u} \in U$ there exists a unique point $(x, y, z) \in V$ as above. Suppose at the point $\mathbf{u}_0 \in U$, there is an infinitesimal box having sides du_1, du_2, du_3 . Then this little box would correspond to something in V . What? Consider the mapping from U to V defined by

$$\mathbf{x} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} f_1(u_1, u_2, u_3) \\ f_2(u_1, u_2, u_3) \\ f_3(u_1, u_2, u_3) \end{pmatrix} = \mathbf{f}(\mathbf{u}) \quad (27.4)$$

which takes a point \mathbf{u} in U and sends it to the point in V which is identified as $(x, y, z)^T \equiv \mathbf{x}$. What happens to a point of the infinitesimal box? Such a point is of the form

$$(u_{01} + s_1 du_1, u_{02} + s_2 du_2, u_{03} + s_3 du_3),$$

where $s_i \geq 0$ and $\sum_i s_i \leq 1$. Also, from the definition of the derivative,

$$\begin{aligned} & \mathbf{f}(u_{10} + s_1 du_1, u_{20} + s_2 du_2, u_{30} + s_3 du_3) - \mathbf{f}(u_{01}, u_{02}, u_{03}) = \\ & D\mathbf{f}(u_{10}, u_{20}, u_{30}) \begin{pmatrix} s_1 du_1 \\ s_2 du_2 \\ s_3 du_3 \end{pmatrix} + o \begin{pmatrix} s_1 du_1 \\ s_2 du_2 \\ s_3 du_3 \end{pmatrix} \end{aligned}$$

where the last term may be taken equal to $\mathbf{0}$ since the vector $(s_1 du_1, s_2 du_2, s_3 du_3)^T$ is infinitesimal, meaning nothing precise, but conveying the idea that it is surpassingly small. Therefore, a point of this infinitesimal box is sent to the vector

$$\begin{aligned} & \overbrace{\left(\frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_1}, \frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_2}, \frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_3} \right)}^{=D\mathbf{f}(u_{10}, u_{20}, u_{30})} \begin{pmatrix} s_1 du_1 \\ s_2 du_2 \\ s_3 du_3 \end{pmatrix} = \\ & s_1 \frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_1} du_1 + s_2 \frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_2} du_2 + s_3 \frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_3} du_3, \end{aligned}$$

a point of the infinitesimal parallelepiped determined by the vectors

$$\left\{ \frac{\partial \mathbf{x}(u_{10}, u_{20}, u_{30})}{\partial u_1} du_1, \frac{\partial \mathbf{x}(u_{10}, u_{20}, u_{30})}{\partial u_2} du_2, \frac{\partial \mathbf{x}(u_{10}, u_{20}, u_{30})}{\partial u_3} du_3 \right\}.$$

The situation is no different for general coordinate systems in any dimension. In general, $\mathbf{x} = \mathbf{f}(\mathbf{u})$ where $\mathbf{u} \in U$, a subset of \mathbb{R}^p and \mathbf{x} is a point in V , a subset of p dimensional space. Thus, letting the Cartesian coordinates of \mathbf{x} be given by $\mathbf{x} = (x_1, \dots, x_p)^T$, each x_i being a function of \mathbf{u} , an infinitesimal box located at \mathbf{u}_0 corresponds to an infinitesimal parallelepiped located at $\mathbf{f}(\mathbf{u}_0)$ which is determined by the p vectors $\left\{ \frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_i} du_i \right\}_{i=1}^p$. From Definition 27.5.1, the volume of this infinitesimal parallelepiped located at $\mathbf{f}(\mathbf{u}_0)$ is given by

$$\left(\det \left(\frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_i} du_i \cdot \frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_j} du_j \right) \right)^{1/2} \quad (27.5)$$

in which there is no sum on the repeated index. As pointed out above, after Definition 27.5.1, if there are p vectors in \mathbb{R}^p , $\{\mathbf{v}_1, \dots, \mathbf{v}_p\}$,

$$\det(\mathbf{v}_i \cdot \mathbf{v}_j)^{1/2} = |\det(\mathbf{v}_1, \dots, \mathbf{v}_p)| \quad (27.6)$$

where this last matrix is the $p \times p$ matrix which has the i^{th} column equal to \mathbf{v}_i . Therefore, from the properties of determinants, 27.5 equals

$$\left| \det \left(\frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_1} du_1, \dots, \frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_p} du_p \right) \right| = \left| \det \left(\frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_1}, \dots, \frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_p} \right) \right| du_1 \cdots du_p$$

This is the infinitesimal chunk of volume corresponding to the point $\mathbf{f}(\mathbf{u}_0)$ in V . The advantage of $\left(\det \left(\frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_i} du_i \cdot \frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_j} du_j \right) \right)^{1/2}$ is that it goes on making sense even if the vectors $\frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_j}$ are in \mathbb{R}^q for $q > p$ thus allowing the consideration of integrals on p dimensional surfaces in \mathbb{R}^q . However, this will not be pursued much further in this book.

Definition 27.5.2 Let $\mathbf{x} = \mathbf{f}(\mathbf{u})$ be as described above. Then the symbol

$$\frac{\partial(x_1, \dots, x_p)}{\partial(u_1, \dots, u_p)},$$

called the Jacobian determinant, is defined by

$$\det \left(\frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_1}, \dots, \frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_p} \right) \equiv \frac{\partial(x_1, \dots, x_p)}{\partial(u_1, \dots, u_p)}.$$

Also, the symbol $\left| \frac{\partial(x_1, \dots, x_p)}{\partial(u_1, \dots, u_p)} \right| du_1 \cdots du_p$ is called the volume element or increment of volume, or increment of area.

This has given motivation for the following fundamental procedure often called the **change of variables formula** which holds under fairly general conditions.

Procedure 27.5.3 Suppose U is an open subset of \mathbb{R}^p for $p > 0$ and suppose $\mathbf{f} : U \rightarrow \mathbf{f}(U)$ is a C^1 function which is one to one, $\mathbf{x} = \mathbf{f}(\mathbf{u})$.² Then if $h : \mathbf{f}(U) \rightarrow \mathbb{R}$, is integrable,

$$\int_U h(\mathbf{f}(\mathbf{u})) \left| \frac{\partial(x_1, \dots, x_p)}{\partial(u_1, \dots, u_p)} \right| dV = \int_{\mathbf{f}(U)} h(\mathbf{x}) dV.$$

Example 27.5.4 Find the area of the region in \mathbb{R}^2 which is determined by the lines $y = 2x$, $y = (1/2)x$, $x + y = 1$, $x + y = 3$.

You might sketch this region. You will find it is an ugly quadrilateral. Let $u = x + y$ and $v = \frac{y}{x}$. The reason for this is that the given region corresponds to $(u, v) \in [1, 3] \times [\frac{1}{2}, 2]$, a nice rectangle. Now we need to solve for x, y to obtain the Jacobian. A little computation shows that

$$x = \frac{u}{v+1}, \quad y = \frac{uv}{v+1}$$

Therefore, $\frac{\partial(x, y)}{\partial(u, v)}$ is

$$\det \begin{pmatrix} \frac{1}{v+1} & -\frac{u}{(v+1)^2} \\ \frac{v}{v+1} & \frac{u}{(v+1)^2} \end{pmatrix} = \frac{u}{(v+1)^2}.$$

Therefore, the area of this quadrilateral is

$$\int_{1/2}^2 \int_1^3 \frac{u}{(v+1)^2} du dv = \frac{4}{3}.$$

27.6 Exercises

1. Verify the three dimensional volume increment in spherical coordinates is

$$\rho^2 \sin(\phi) d\rho d\phi d\theta.$$

2. Find the area of the bounded region R , determined by $5x + y = 1$, $5x + y = 9$, $y = 2x$, and $y = 5x$.
3. Find the area of the bounded region R , determined by $y + 2x = 6$, $y + 2x = 10$, $y = 3x$, and $y = 4x$.

²This will cause non overlapping infinitesimal boxes in U to be mapped to non overlapping infinitesimal parallelepipeds in V .

Also, in the context of the Riemann integral we should say more about the set U in any case the function h . These conditions are mainly technical however, and since a mathematically respectable treatment will not be attempted for this theorem in this part of the book, I think it best to give a memorable version of it which is essentially correct in all examples of interest. The simple statement above is just fine if you are using the Lebesgue integral. This integral and a slightly less general theorem is proved in a special chapter on the integral.

4. A solid, R is determined by $3x + y = 2$, $3x + y = 4$, $y = x$, and $y = 2x$ and the density is $\rho = x$. Find the total mass of R .
5. A solid, R is determined by $4x + 2y = 1$, $4x + 2y = 9$, $y = x$, and $y = 6x$ and the density is $\rho = y$. Find the total mass of R .
6. A solid, R is determined by $3x + y = 3$, $3x + y = 10$, $y = 3x$, and $y = 5x$ and the density is $\rho = y^{-1}$. Find the total mass of R .
7. Find a 2×2 matrix A which maps the equilateral triangle having vertices at

$$(0, 0), (1, 0), \text{ and } \left(1/2, \sqrt{3}/2\right)$$

to the triangle having vertices at $(0, 0)$, (a, b) , and (c, d) where (c, d) is not a multiple of (a, b) . Find the area of this last triangle by using the cross product. Next find the area of this triangle using the change of variables formula and the fact that the area of the equilateral triangle is $\frac{\sqrt{3}}{4}$.

8. Find the volume of the region E , bounded by the ellipsoid, $\frac{1}{4}x^2 + y^2 + z^2 = 1$.
9. Here are three vectors. $(4, 1, 2)^T$, $(5, 0, 2)^T$, and $(3, 1, 3)^T$. These vectors determine a parallelepiped, R , which is occupied by a solid having density $\rho = x$. Find the mass of this solid.
10. Here are three vectors. $(5, 1, 6)^T$, $(6, 0, 6)^T$, and $(4, 1, 7)^T$. These vectors determine a parallelepiped, R , which is occupied by a solid having density $\rho = y$. Find the mass of this solid.
11. Here are three vectors. $(5, 2, 9)^T$, $(6, 1, 9)^T$, and $(4, 2, 10)^T$. These vectors determine a parallelepiped, R , which is occupied by a solid having density $\rho = y + x$. Find the mass of this solid.
12. Compute the volume of a sphere of radius R using cylindrical coordinates.
13. Fill in all details for the following argument that

$$\int_0^\infty e^{-x^2} dx = \frac{1}{2}\sqrt{\pi}.$$

Let $I = \int_0^\infty e^{-x^2} dx$. Then

$$I^2 = \int_0^\infty \int_0^\infty e^{-(x^2+y^2)} dx dy = \int_0^{\pi/2} \int_0^\infty r e^{-r^2} dr d\theta = \frac{1}{4}\pi$$

from which the result follows.

14. 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.
15. Show using Problem 13 that $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$. Recall $\Gamma(\alpha) \equiv \int_0^\infty e^{-t} t^{\alpha-1} dt$.
16. Let $p, q > 0$ and define $B(p, q) = \int_0^1 x^{p-1} (1-x)^{q-1} dx$. 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.

27.7 The Moment of Inertia and Center of Mass

The methods used to evaluate multiple integrals make possible the determination of centers of mass and moments of inertia for solids. This leads to the following definition.

Definition 27.7.1 *Let a solid occupy a region R such that its density is $\rho(\mathbf{x})$ for \mathbf{x} a point in R and let L be a line. For $\mathbf{x} \in R$, let $l(\mathbf{x})$ be the distance from the point \mathbf{x} to the line L . The moment of inertia of the solid is defined as*

$$I = \int_R l(\mathbf{x})^2 \rho(\mathbf{x}) dV.$$

Letting $(\bar{x}, \bar{y}, \bar{z})$ denote the Cartesian coordinates of the center of mass,

$$\bar{x} = \frac{\int_R x \rho(\mathbf{x}) dV}{\int_R \rho(\mathbf{x}) dV}, \quad \bar{y} = \frac{\int_R y \rho(\mathbf{x}) dV}{\int_R \rho(\mathbf{x}) dV}, \quad \bar{z} = \frac{\int_R z \rho(\mathbf{x}) dV}{\int_R \rho(\mathbf{x}) dV}$$

where x, y, z are the Cartesian coordinates of the point at \mathbf{x} .

The reason the moment of inertia is of interest has to do with the total kinetic energy of a solid occupying the region R which is rotating about the line L . Suppose its angular velocity is ω . Then the kinetic energy of an infinitesimal chunk of volume located at point \mathbf{x} is $\frac{1}{2} \rho(\mathbf{x}) (l(\mathbf{x}) \omega)^2 dV$. Then using an integral to add these up, it follows the total kinetic energy is

$$\frac{1}{2} \int_R \rho(\mathbf{x}) l(\mathbf{x})^2 dV \omega^2 = \frac{1}{2} I \omega^2$$

Thus in the consideration of a rotating body, the moment of inertia takes the place of mass when angular velocity takes the place of speed.

As to the center of mass, its significance is that it gives the point at which the mass will balance. To see this presented in terms of point masses, see Definition 14.5.4. Here the sums are replaced with integrals.

Example 27.7.2 *Let a solid occupy the three dimensional region R and suppose the density is ρ . What is the moment of inertia of this solid about the z axis? What is the center of mass?*

Here the little masses would be of the form $\rho(\mathbf{x}) dV$ where \mathbf{x} is a point of R . Therefore, the contribution of this mass to the moment of inertia would be $(x^2 + y^2) \rho(\mathbf{x}) dV$ where the Cartesian coordinates of the point \mathbf{x} are (x, y, z) . Then summing these up as an integral, yields the following for the moment of inertia.

$$\int_R (x^2 + y^2) \rho(\mathbf{x}) dV. \quad (27.7)$$

To find the center of mass, sum up $\mathbf{r} \rho dV$ for the points in R and divide by the total mass. In Cartesian coordinates, where $\mathbf{r} = (x, y, z)$, this means to sum up vectors of the form $(x \rho dV, y \rho dV, z \rho dV)$ and divide by the total mass. Thus the Cartesian coordinates of the center of mass are

$$\left(\frac{\int_R x \rho dV}{\int_R \rho dV}, \frac{\int_R y \rho dV}{\int_R \rho dV}, \frac{\int_R z \rho dV}{\int_R \rho dV} \right) \equiv \frac{\int_R \mathbf{r} \rho dV}{\int_R \rho dV}.$$

Here is a specific example.

Example 27.7.3 Find the moment of inertia about the z axis and center of mass of the solid which occupies the region R defined by $9 - (x^2 + y^2) \geq z \geq 0$ if the density is $\rho(x, y, z) = \sqrt{x^2 + y^2}$.

This moment of inertia is $\int_R (x^2 + y^2) \sqrt{x^2 + y^2} dV$ and the easiest way to find this integral is to use cylindrical coordinates. Thus the answer is

$$\int_0^{2\pi} \int_0^3 \int_0^{9-r^2} r^3 r dz dr d\theta = \frac{8748}{35} \pi.$$

To find the center of mass, note the x and y coordinates of the center of mass,

$$\frac{\int_R x \rho dV}{\int_R \rho dV}, \frac{\int_R y \rho dV}{\int_R \rho dV}$$

both equal zero because the above shape is symmetric about the z axis and ρ is also symmetric in its values. Thus $x \rho dV$ will cancel with $-x \rho dV$ and a similar conclusion will hold for the y coordinate. It only remains to find the z coordinate of the center of mass, \bar{z} . In polar coordinates, $\rho = r$ and so,

$$\bar{z} = \frac{\int_R z \rho dV}{\int_R \rho dV} = \frac{\int_0^{2\pi} \int_0^3 \int_0^{9-r^2} z r^2 dz dr d\theta}{\int_0^{2\pi} \int_0^3 \int_0^{9-r^2} r^2 dz dr d\theta} = \frac{18}{7}.$$

Thus the center of mass will be $(0, 0, \frac{18}{7})$.

27.8 Exercises

1. Let R denote the finite region bounded by $z = 4 - x^2 - y^2$ and the xy plane. Find z_c , the z coordinate of the center of mass if the density σ is a constant.
2. Let R denote the finite region bounded by $z = 4 - x^2 - y^2$ and the xy plane. Find z_c , the z coordinate of the center of mass if the density σ is equals $\sigma(x, y, z) = z$.
3. Find the mass and center of mass of the region between the surfaces $z = -y^2 + 8$ and $z = 2x^2 + y^2$ if the density equals $\sigma = 1$.
4. Find the mass and center of mass of the region between the surfaces $z = -y^2 + 8$ and $z = 2x^2 + y^2$ if the density equals $\sigma(x, y, z) = x^2$.
5. The two cylinders, $x^2 + y^2 = 4$ and $y^2 + z^2 = 4$ intersect in a region R . Find the mass and center of mass if the density σ , is given by $\sigma(x, y, z) = z^2$.
6. The two cylinders, $x^2 + y^2 = 4$ and $y^2 + z^2 = 4$ intersect in a region R . Find the mass and center of mass if the density σ , is given by $\sigma(x, y, z) = 4 + z$.
7. Find the mass and center of mass of the set (x, y, z) such that $\frac{x^2}{4} + \frac{y^2}{9} + z^2 \leq 1$ if the density is $\sigma(x, y, z) = 4 + y + z$.
8. Let R denote the finite region bounded by $z = 9 - x^2 - y^2$ and the xy plane. Find the moment of inertia of this shape about the z axis given the density equals 1.

9. Let R denote the finite region bounded by $z = 9 - x^2 - y^2$ and the xy plane. Find the moment of inertia of this shape about the x axis given the density equals 1.
10. Let B be a solid ball of constant density and radius R . Find the moment of inertia about a line through a diameter of the ball. You should get $\frac{2}{5}R^2M$ where M is the mass..
11. Let B be a solid ball of density $\sigma = \rho$ where ρ is the distance to the center of the ball which has radius R . Find the moment of inertia about a line through a diameter of the ball. Write your answer in terms of the total mass and the radius as was done in the constant density case.
12. Let C be a solid cylinder of constant density and radius R . Find the moment of inertia about the axis of the cylinder
You should get $\frac{1}{2}R^2M$ where M is the mass.
13. Let C be a solid cylinder of constant density and radius R and mass M and let B be a solid ball of radius R and mass M . The cylinder and the ball are placed on the top of an inclined plane and allowed to roll to the bottom. Which one will arrive first and why?
14. A ball of radius 4 has a cone taken out of the top which has an angle of $\pi/2$ and then a cone taken out of the bottom which has an angle of $\pi/3$. If the density is $\lambda = \rho$, find the z component of the center of mass.
15. A ball of radius 4 has a cone taken out of the top which has an angle of $\pi/2$ and then a cone taken out of the bottom which has an angle of $\pi/3$. If the density is $\lambda = \rho$, find the moment of inertia about the z axis.
16. Suppose a solid of mass M occupying the region B has moment of inertia, I_l about a line, l which passes through the center of mass of M and let l_1 be another line parallel to l and at a distance of a from l . Then the parallel axis theorem states $I_{l_1} = I_l + a^2M$. Prove the parallel axis theorem. **Hint:** Choose axes such that the z axis is l and l_1 passes through the point $(a, 0)$ in the xy plane.
17. * Using the parallel axis theorem find the moment of inertia of a solid ball of radius R and mass M about an axis located at a distance of a from the center of the ball. Your answer should be $Ma^2 + \frac{2}{5}MR^2$.
18. Consider all axes in computing the moment of inertia of a solid. Will the smallest possible moment of inertia always result from using an axis which goes through the center of mass?
19. Find the moment of inertia of a solid thin rod of length l , mass M , and constant density about an axis through the center of the rod perpendicular to the axis of the rod. You should get $\frac{1}{12}l^2M$.
20. Using the parallel axis theorem, find the moment of inertia of a solid thin rod of length l , mass M , and constant density about an axis through an end of the rod perpendicular to the axis of the rod. You should get $\frac{1}{3}l^2M$.

21. Let the angle between the z axis and the sides of a right circular cone be α . Also assume the height of this cone is h . Find the z coordinate of the center of mass of this cone in terms of α and h assuming the density is constant.
22. Let the angle between the z axis and the sides of a right circular cone be α . Also assume the height of this cone is h . Assuming the density is $\sigma = 1$, find the moment of inertia about the z axis in terms of α and h .
23. Let R denote the part of the solid ball, $x^2 + y^2 + z^2 \leq R^2$ which lies in the first octant. That is $x, y, z \geq 0$. Find the coordinates of the center of mass if the density is constant. Your answer for one of the coordinates for the center of mass should be $(3/8)R$.
24. Show that in general for \mathbf{L} angular momentum, $\frac{d\mathbf{L}}{dt} = \mathbf{\Gamma}$ where $\mathbf{\Gamma}$ is the total torque, $\mathbf{\Gamma} \equiv \sum \mathbf{r}_i \times \mathbf{F}_i$ where \mathbf{F}_i is the force on the i^{th} point mass.

Chapter 28

The Integral on Two Dimensional Surfaces in \mathbb{R}^3

A parametric surface is the image of a vector valued function of two variables. Earlier, vector valued functions of one variable were considered in the study of space curves. Here there are two independent variables. This is why the result could be expected to be a surface. For example, you could have

$$\mathbf{r}(s,t) = \begin{pmatrix} x & y & z \end{pmatrix} = \begin{pmatrix} s+t & \cos(s)\sin(s) & ts \end{pmatrix}$$

for $(s,t) \in (0,1) \times (0,1)$. Each value of (s,t) gives a point on this surface. The surface is smooth if all the component functions are C^1 and $\mathbf{r}_s \times \mathbf{r}_t(s,t) \neq 0$. This last condition assures the existence of a well defined normal vector to the surface, namely $\mathbf{r}_s \times \mathbf{r}_t(s,t)$. Recall from the material on space curves that $\mathbf{r}_t, \mathbf{r}_s$ are both tangent to curves which lie in this surface. If this cross product were 0, you would get points or creases in the surface.

28.1 The Two Dimensional Area in \mathbb{R}^3

Consider a function defined on a two dimensional surface. Imagine taking the value of this function at a point, multiplying this value by the area of an infinitesimal chunk of area located at this point and then adding these together. The only difference is that now you need a two dimensional chunk of area rather than one dimensional.

Definition 28.1.1 Let $\mathbf{u}_1, \mathbf{u}_2$ be vectors in \mathbb{R}^3 . The 2 dimensional parallelogram determined by these vectors will be denoted by $P(\mathbf{u}_1, \mathbf{u}_2)$ and it is defined as

$$P(\mathbf{u}_1, \mathbf{u}_2) \equiv \left\{ \sum_{j=1}^2 s_j \mathbf{u}_j : s_j \in [0,1] \right\}.$$

Then the area of this parallelogram is

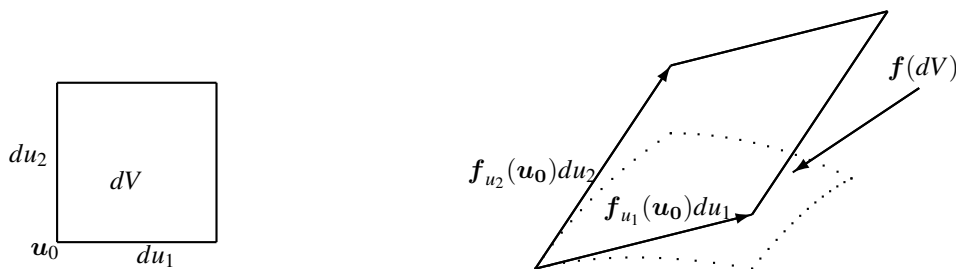
$$\text{area } P(\mathbf{u}_1, \mathbf{u}_2) \equiv |\mathbf{u}_1 \times \mathbf{u}_2| = \det(G)^{1/2}$$

where $G_{ij} \equiv \mathbf{u}_i \cdot \mathbf{u}_j$.

To see the last claim,

$$\begin{aligned} |\mathbf{u}_1 \times \mathbf{u}_2|^2 &= |\mathbf{u}_1|^2 |\mathbf{u}_2|^2 \sin^2(\theta) = |\mathbf{u}_1|^2 |\mathbf{u}_2|^2 (1 - \cos^2(\theta)) \\ &= |\mathbf{u}_1|^2 |\mathbf{u}_2|^2 - (\mathbf{u}_1 \cdot \mathbf{u}_2)^2 = \det(G)^2 \end{aligned}$$

Suppose then that $\mathbf{x} = \mathbf{f}(\mathbf{u})$ where $\mathbf{u} \in U$, a subset of \mathbb{R}^2 and \mathbf{x} is a point in V , a subset of 3 dimensional space. Thus, letting the Cartesian coordinates of \mathbf{x} be given by $\mathbf{x} = (x_1, x_2, x_3)^T$, each x_i being a function of \mathbf{u} , an infinitesimal rectangle located at \mathbf{u}_0 corresponds to an infinitesimal parallelogram located at $\mathbf{f}(\mathbf{u}_0)$ which is determined by the 2 vectors $\left\{ \frac{\partial \mathbf{f}(\mathbf{u}_0)}{\partial u_i} du_i \right\}_{i=1}^2$, each of which is tangent to the surface defined by $\mathbf{x} = \mathbf{f}(\mathbf{u})$. (No sum on the repeated index.)



From Definition 28.1.1, the two dimensional volume of this infinitesimal parallelepiped located at $\mathbf{f}(\mathbf{u}_0)$ is given by

$$\left| \frac{\partial \mathbf{f}(\mathbf{u}_0)}{\partial u_1} du_1 \times \frac{\partial \mathbf{f}(\mathbf{u}_0)}{\partial u_2} du_2 \right| = \left| \frac{\partial \mathbf{f}(\mathbf{u}_0)}{\partial u_1} \times \frac{\partial \mathbf{f}(\mathbf{u}_0)}{\partial u_2} \right| du_1 du_2 \quad (28.1)$$

$$= |\mathbf{f}_{u_1} \times \mathbf{f}_{u_2}| du_1 du_2 \quad (28.2)$$

It might help to think of a lizard. The infinitesimal parallelepiped is like a very small scale on a lizard. This is the essence of the idea. To define the area of the lizard sum up areas of individual scales¹. If the scales are small enough, their sum would serve as a good approximation to the area of the lizard.



This motivates the following fundamental procedure which I hope is extremely familiar from the earlier material.

¹This beautiful lizard is a *Sceloporus magister*. It was photographed by C. Riley Nelson who is in the Zoology department at Brigham Young University © 2004 in Kane Co. Utah. The lizard is a little less than one foot in length.

Procedure 28.1.2 Suppose U is a subset of \mathbb{R}^2 and suppose $\mathbf{f} : U \rightarrow \mathbf{f}(U) \subseteq \mathbb{R}^3$ is a one to one and C^1 function. Then if $h : \mathbf{f}(U) \rightarrow \mathbb{R}$, define the 2 dimensional surface integral $\int_{\mathbf{f}(U)} h(\mathbf{x}) dA$ according to the following formula.

$$\begin{aligned} \int_{\mathbf{f}(U)} h(\mathbf{x}) dA &\equiv \int_U h(\mathbf{f}(\mathbf{u})) |\mathbf{f}_{u_1}(\mathbf{u}) \times \mathbf{f}_{u_2}(\mathbf{u})| du_1 du_2 \\ &= \int_U h(\mathbf{f}(\mathbf{u})) \det(G(\mathbf{u}))^{1/2} du_1 du_2 \end{aligned}$$

$$\text{where } G(\mathbf{u}) = \begin{pmatrix} \mathbf{f}_{u_1} \cdot \mathbf{f}_{u_1} & \mathbf{f}_{u_1} \cdot \mathbf{f}_{u_2} \\ \mathbf{f}_{u_2} \cdot \mathbf{f}_{u_1} & \mathbf{f}_{u_2} \cdot \mathbf{f}_{u_2} \end{pmatrix}.$$

Note that the Jacobian for change of variables and the Jacobian to be used in surface integrals are really both special cases of the same general theory involving the square root of the determinant of the matrix G . This matrix is called the metric tensor and will be considered more later.

Definition 28.1.3 It is customary to write $|\mathbf{f}_{u_1}(\mathbf{u}) \times \mathbf{f}_{u_2}(\mathbf{u})| = \frac{\partial(x_1, x_2, x_3)}{\partial(u_1, u_2)}$ because this new notation generalizes to far more general situations for which the cross product is not defined. For example, one can consider three dimensional surfaces in \mathbb{R}^8 .

Example 28.1.4 Consider the surface given by $z = x^2$ for $(x, y) \in [0, 1] \times [0, 1] = U$. Find the surface area of this surface.

The first step in using the above is to write this surface in the form $\mathbf{x} = \mathbf{f}(\mathbf{u})$. This is easy to do if you let $\mathbf{u} = (x, y)$. Then $\mathbf{f}(x, y) = (x, y, x^2)$. If you like, let $x = u_1$ and $y = u_2$. What is $\frac{\partial(x_1, x_2, x_3)}{\partial(x, y)} = |\mathbf{f}_x \times \mathbf{f}_y|$?

$$\mathbf{f}_x = \begin{pmatrix} 1 & 0 & 2x \end{pmatrix}^T, \mathbf{f}_y = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}^T$$

and so

$$|\mathbf{f}_x \times \mathbf{f}_y| = \left| \begin{pmatrix} 1 & 0 & 2x \end{pmatrix}^T \times \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}^T \right| = \sqrt{1 + 4x^2}$$

and so the area element is $\sqrt{1 + 4x^2} dx dy$ and the surface area is obtained by integrating the function $h(\mathbf{x}) \equiv 1$. Therefore, this area is

$$\int_{\mathbf{f}(U)} dA = \int_0^1 \int_0^1 \sqrt{1 + 4x^2} dx dy = \frac{1}{2} \sqrt{5} - \frac{1}{4} \ln(-2 + \sqrt{5})$$

which can be obtained by using the trig. substitution, $2x = \tan \theta$ on the inside integral.

Note this all depends on being able to write the surface in the form, $\mathbf{x} = \mathbf{f}(\mathbf{u})$ for $\mathbf{u} \in U \subseteq \mathbb{R}^p$. Surfaces obtained in this form are called parametrically defined surfaces. These are best but sometimes you have some other description of a surface and in these cases things can get pretty intractable. For example, you might have a level surface of the form $3x^2 + 4y^4 + z^6 = 10$. In this case, you could solve for z using methods of algebra. Thus $z = \sqrt[6]{10 - 3x^2 - 4y^4}$ and a parametric description of part of this level surface is $(x, y, \sqrt[6]{10 - 3x^2 - 4y^4})$ for $(x, y) \in U$ where $U = \{(x, y) : 3x^2 + 4y^4 \leq 10\}$. But what if the level surface was something like

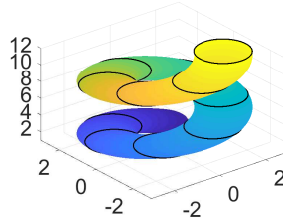
$$\sin(x^2 + \ln(7 + y^2 \sin x)) + \sin(zy) e^z = 11 \sin(xyz)?$$

I really do not see how to use methods of algebra to solve for some variable in terms of the others. It isn't even clear to me whether there are any points $(x, y, z) \in \mathbb{R}^3$ satisfying this particular relation. However, if a point satisfying this relation can be identified, the implicit function theorem from advanced calculus can usually be used to assert one of the variables is a function of the others, proving the existence of a parametrization at least locally. The problem is, this theorem does not give the answer in terms of known functions so this is not much help. Finding a parametric description of a surface is a hard problem and there are no easy answers. This is a good example which illustrates the gulf between theory and practice.

Example 28.1.5 Let $U = [0, 12] \times [0, 2\pi]$ and let $\mathbf{f} : U \rightarrow \mathbb{R}^3$ be given by

$$\mathbf{f}(t, s) \equiv (2 \cos t + \cos s, 2 \sin t + \sin s, t)^T$$

Find a double integral for the surface area. A graph of this surface is drawn below.



Then $\mathbf{f}_t = (-2 \sin t \quad 2 \cos t \quad 1)^T$, $\mathbf{f}_s = (-\sin s \quad \cos s \quad 0)^T$ and

$$\mathbf{f}_t \times \mathbf{f}_s = \begin{pmatrix} -\cos s \\ -\sin s \\ -2 \sin t \cos s + 2 \cos t \sin s \end{pmatrix}$$

and so $\frac{\partial(x_1, x_2, x_3)}{\partial(t, s)} =$

$$|\mathbf{f}_t \times \mathbf{f}_s| = \sqrt{5 - 4 \sin^2 t \sin^2 s - 8 \sin t \sin s \cos t \cos s - 4 \cos^2 t \cos^2 s}.$$

Therefore, the desired integral giving the area is

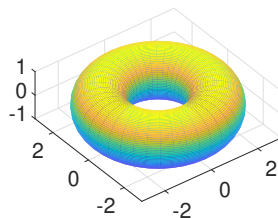
$$\int_0^{2\pi} \int_0^{12} \sqrt{5 - 4 \sin^2 t \sin^2 s - 8 \sin t \sin s \cos t \cos s - 4 \cos^2 t \cos^2 s} dt ds.$$

If you really needed to find the number this equals, how would you go about finding it? This is an interesting question and there is no single right answer. You should think about this. Here is an example for which you will be able to find the integrals.

Example 28.1.6 Let $U = [0, 2\pi] \times [0, 2\pi]$ and for $(t, s) \in U$, let

$$\mathbf{f}(t, s) = (2 \cos t + \cos s, -2 \sin t - \sin s, \sin s)^T.$$

Find the area of $\mathbf{f}(U)$. This is the surface of a donut shown below. The fancy name for this shape is a torus.



To find its area,

$$\mathbf{f}_t = \begin{pmatrix} -2 \sin t - \sin t \cos s \\ -2 \cos t - \cos t \cos s \\ 0 \end{pmatrix}, \mathbf{f}_s = \begin{pmatrix} -\cos t \sin s \\ \sin t \sin s \\ \cos s \end{pmatrix}$$

and so $|\mathbf{f}_t \times \mathbf{f}_s| = (\cos s + 2)$ so the area element is $(\cos s + 2) ds dt$ and the area is

$$\int_0^{2\pi} \int_0^{2\pi} (\cos s + 2) ds dt = 8\pi^2$$

Example 28.1.7 Let $U = [0, 2\pi] \times [0, 2\pi]$ and for $(t, s) \in U$, let

$$\mathbf{f}(t, s) = (2 \cos t + \cos t \cos s, -2 \sin t - \sin t \cos s, \sin s)^T.$$

Find $\int_{\mathbf{f}(U)} h dV$ where $h(x, y, z) = x^2$.

Everything is the same as the preceding example except this time it is an integral of a function. The area element is $(\cos s + 2) ds dt$ and so the integral called for is

$$\int_{\mathbf{f}(U)} h dA = \int_0^{2\pi} \int_0^{2\pi} \left(\overbrace{2 \cos t + \cos t \cos s}^{x \text{ on the surface}} \right)^2 (\cos s + 2) ds dt = 22\pi^2$$

28.2 Surfaces of the Form $z = f(x, y)$

The special case where a surface is in the form $z = f(x, y)$, $(x, y) \in U$, yields a simple formula which is used most often in this situation. You write the surface parametrically in the form $\mathbf{f}(x, y) = (x, y, f(x, y))^T$ such that $(x, y) \in U$. Then

$$\mathbf{f}_x = \begin{pmatrix} 1 & 0 & f_x \end{pmatrix}^T, \mathbf{f}_y = \begin{pmatrix} 0 & 1 & f_y \end{pmatrix}^T$$

and $|\mathbf{f}_x \times \mathbf{f}_y| = \sqrt{1 + f_y^2 + f_x^2}$ so the area element is

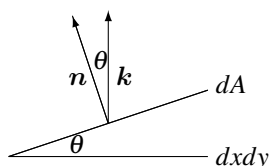
$$\sqrt{1 + f_y^2 + f_x^2} dx dy.$$

When the surface of interest comes in this simple form, people generally use this area element directly rather than worrying about a parametrization and taking cross products.

In the case where the surface is of the form $x = f(y, z)$ for $(y, z) \in U$, the area element is obtained similarly and is $\sqrt{1 + f_y^2 + f_z^2} dy dz$. I think you can guess what the area element is if $y = f(x, z)$. It also generalizes immediately to higher dimensions where $x_{k_i} = f(x_1, \dots, x_{k_i-1}, x_{k_i+1}, \dots, x_n)$.

There is also a simple geometric description of these area elements. Consider the surface $z = f(x, y)$. This is a level surface of the function of three variables $z - f(x, y)$. In

fact the surface is simply $z - f(x, y) = 0$. Now consider the gradient of this function of three variables. The gradient is perpendicular to the surface and the third component is positive in this case. This gradient is $(-f_x, -f_y, 1)$ and so the unit upward normal is just $\frac{1}{\sqrt{1+f_x^2+f_y^2}}(-f_x, -f_y, 1)$. Now consider the following picture.



In this picture, you are looking at a chunk of area on the surface seen on edge and so it seems reasonable to expect to have $dx dy = dA \cos \theta$. But it is easy to find $\cos \theta$ from the picture and the properties of the dot product.

$$\cos \theta = \frac{\mathbf{n} \cdot \mathbf{k}}{|\mathbf{n}| |\mathbf{k}|} = \frac{1}{\sqrt{1+f_x^2+f_y^2}}.$$

Therefore, $dA = \sqrt{1+f_x^2+f_y^2} dx dy$ as claimed.

Example 28.2.1 Let $z = \sqrt{x^2+y^2}$ where $(x, y) \in U$ for

$$U = \{(x, y) : x^2 + y^2 \leq 4\}$$

Find $\int_S h dS$ where $h(x, y, z) = x + z$ and S is the surface whose parametrical description is $(x, y, \sqrt{x^2+y^2})$ for $(x, y) \in U$.

Here you can see directly the angle in the above picture is $\frac{\pi}{4}$ and so $dA = \sqrt{2} dx dy$. If you do not see this or if it is unclear, simply compute $\sqrt{1+f_x^2+f_y^2}$ and you will find it is $\sqrt{2}$. Therefore, using polar coordinates,

$$\begin{aligned} \int_S h dS &= \int_U (x + \sqrt{x^2+y^2}) \sqrt{2} dA \\ &= \sqrt{2} \int_0^{2\pi} \int_0^2 (r \cos \theta + r) r dr d\theta = \frac{16}{3} \sqrt{2} \pi. \end{aligned}$$

I have been purposely vague about precise mathematical conditions necessary for the above procedures. This is because the precise mathematical conditions which are usually cited are very technical and at the same time far too restrictive. The most general conditions under which these sorts of procedures are valid include things like Lipschitz functions defined on very general sets. These are functions satisfying a Lipschitz condition of the form $|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| \leq K |\mathbf{x} - \mathbf{y}|$. For example, $y = |x|$ is Lipschitz continuous. This function does not have a derivative at every point. So it is with Lipschitz functions. However, it turns out these functions have derivatives at enough points to push everything through but this requires considerations involving the Lebesgue integral.

28.3 MATLAB and Graphing Surfaces

I will illustrate with an example.

```
[s,t]=meshgrid(0:.02*pi:2*pi,0:.02*pi:pi);
[u,v]=meshgrid(0:.02*pi:2*pi,-1.4:.2:1.4);
hold on
surf(sin(t).*cos(s),sin(t).*sin(s),cos(t),'edgecolor','none')
alpha .7
surf(.5*cos(u),.5*sin(u),v,'edgecolor','none')
axis equal
```

This graphs two surfaces, a cylinder and a sphere. The .7 makes the sphere slightly transparent. You can adjust this number to be anything between 0 and 1 depending on how transparent you want it to be. If you just wanted to graph the sphere, you could forget about the hold on and simply include the first of the two lines beginning with “surf”. You should experiment with this. These are parametrically defined surfaces because this is more general than a surface of the form $z = f(x,y)$ and the integral is defined on these more general kinds of surfaces. Click on the little curvy arrow on the top of the picture which appears to allow rotating the graph to see it from different angles.

28.4 Piecewise Defined Surfaces

As with curves, you might piece together surfaces. In this section is considered what happens on the place where the two surfaces intersect. First of all, we really don't know how to find the Riemann integral over arbitrary regions. We need to have the region be cylindrical in either the u or the v direction. That is, $u \in [a, b]$ and for each u , the variable v is between $T(u)$ and $B(u)$. Alternatively, $v \in [c, d]$ and for each v , the variable u is between $L(v)$ and $R(v)$ where $L(v) \leq R(v)$. So what is meant by a piecewise smooth surface? Let

$$S \equiv S_1 \cup S_2 \cup \cdots \cup S_m$$

where $S_k \equiv \mathbf{r}_k(D_k)$ where D_k is one of the special regions just described and \mathbf{r}_k is one to one and C^1 on an open set $U_k \supseteq D_k$ such that $\mathbf{r}_u \times \mathbf{r}_v \neq \mathbf{0}$. Then we assume that either $S_k \cap S_j = \emptyset$ or their intersection is $\mathbf{r}_k(l_k) = \mathbf{r}_j(l_j)$ where l_k, l_j are one of the four edges of D_k and D_j respectively. For example, say

$$D_k = \{u \in [a, b], v \in [B(u), T(u)]\}$$

and say l_k is the top edge of D_k , $\{(u, T(u)) : u \in [a, b]\}$. Then from the definition, if f is defined on S , and is 0 off $S_k \cap S_j$,

$$\int_S f dS = \int_a^b \int_{T(u)}^{B(u)} f(u, v) |\mathbf{r}_{ku} \times \mathbf{r}_{kv}| dv du = 0$$

Other situations are exactly similar. The point is, when you have a surface which is defined piecewise as just described, you don't need to bother with the curves of intersection because the two dimensional iterated integral will be zero on these curves. The term for this situation in the context of the Lebesgue integral is that the curve has measure zero. In examples of interest, the situation is usually that surfaces intersect in sets of measure zero and so as far as the integral is concerned, these intersections are irrelevant.

28.5 Flux Integrals

These will be important in the next chapter. The idea is this. You have a surface S and a field of unit normal vectors \mathbf{n} on S . That is, for each point of S there exists a unit normal except for finitely many curves of measure zero. There is also a vector field \mathbf{F} and you want to find $\int_S \mathbf{F} \cdot \mathbf{n} dS$. There is really nothing new here. You just need to compute the function $\mathbf{F} \cdot \mathbf{n}$ and then integrate it over the surface. Here is an example.

Example 28.5.1 Let $\mathbf{F}(x, y, z) = (x, x + z, y)$ and let S be the hemisphere $x^2 + y^2 + z^2 = 4, z \geq 0$. Let \mathbf{n} be the unit normal to S which has nonnegative z component. Find $\int_S \mathbf{F} \cdot \mathbf{n} dS$.

First find the function $\mathbf{F} \cdot \mathbf{n} \equiv (x, x + z, y) \cdot \overbrace{(x, y, z)}^{=\mathbf{n}} \frac{1}{2} = \frac{1}{2}x^2 + \frac{1}{2}(x + z)y + \frac{1}{2}yz$. This follows because the normal is of the form $(2x, 2y, 2z)$ and then when you divide by its length using the fact that $x^2 + y^2 + z^2 = 4$, you obtain that $\mathbf{n} = (x, y, z) \frac{1}{2}$ as claimed. Next it remains to choose a coordinate system for the surface and then to compute the integral. A parametrization is

$$x = 2 \sin \phi \cos \theta, \quad y = 2 \sin \phi \sin \theta, \quad z = 2 \cos \phi$$

and the increment of surface area is then

$$\begin{aligned} & \left| \begin{pmatrix} -2 \sin \phi \sin \theta \\ 2 \sin \phi \cos \theta \\ 0 \end{pmatrix} \times \begin{pmatrix} 2 \cos \phi \cos \theta \\ 2 \cos \phi \sin \theta \\ -2 \sin \phi \end{pmatrix} \right| d\theta d\phi \\ &= \left| \begin{pmatrix} -4 \sin^2 \phi \cos \theta \\ -4 \sin^2 \phi \sin \theta \\ -4 \sin \phi \cos \phi \end{pmatrix} \right| d\theta d\phi = 4 \sin \phi d\theta d\phi \end{aligned}$$

Therefore, since the hemisphere corresponds to $\theta \in [0, 2\pi]$ and $\phi \in [0, \pi/2]$, the integral to work is

$$\begin{aligned} & \int_0^{2\pi} \int_0^{\pi/2} \left[\frac{1}{2} (2 \sin \phi \cos \theta)^2 + \left(\frac{1}{2} (2 \sin \phi \cos \theta + 2 \cos \phi) \right) \right. \\ & \left. (2 \sin \phi \sin \theta) + \frac{1}{2} (2 \sin \phi \sin \theta) 2 \cos \phi \right] 4 \sin \phi d\phi d\theta \end{aligned}$$

Doing the integration, this reduces to $\frac{16}{3}\pi$.

The important thing to notice is that there is no new mathematics here. That which is new is the significance of a flux integral which will be discussed more in the next chapter. In short, this integral often has the interpretation of a measure of how fast something is crossing a surface.

28.6 Exercises

1. Find a parametrization for the intersection of the planes $4x + 2y + 4z = 3$ and $6x - 2y = -1$.

2. Find a parametrization for the intersection of the plane $3x + y + z = 1$ and the circular cylinder $x^2 + y^2 = 1$.
3. Find a parametrization for the intersection of the plane $3x + 2y + 4z = 4$ and the elliptic cylinder $x^2 + 4z^2 = 16$.
4. Find a parametrization for the straight line joining $(1, 3, 1)$ and $(-2, 5, 3)$.
5. Find a parametrization for the intersection of the surfaces $4y + 3z = 3x^2 + 2$ and $3y + 2z = -x + 3$.
6. Find the area of S if S is the part of the circular cylinder $x^2 + y^2 = 4$ which lies between $z = 0$ and $z = 2 + y$.
7. Find the area of S if S is the part of the cone $x^2 + y^2 = 16z^2$ between $z = 0$ and $z = h$.
8. Parametrizing the cylinder $x^2 + y^2 = a^2$ by $x = a \cos v, y = a \sin v, z = u$, show that the area element is $dA = a du dv$.
9. Find the area enclosed by the limaçon $r = 2 + \cos \theta$.
10. Find the surface area of the paraboloid $z = h(1 - x^2 - y^2)$ between $z = 0$ and $z = h$. Take a limit of this area as h decreases to 0.
11. Evaluate $\int_S (1 + x) dA$ where S is the part of the plane $4x + y + 3z = 12$ which is in the first octant.
12. Evaluate $\int_S (1 + x) dA$ where S is the part of the cylinder $x^2 + y^2 = 9$ between $z = 0$ and $z = h$.
13. Evaluate $\int_S (1 + x) dA$ where S is the hemisphere $x^2 + y^2 + z^2 = 4$ between $x = 0$ and $x = 2$.
14. For $(\theta, \alpha) \in [0, 2\pi] \times [0, 2\pi]$, let

$$\mathbf{f}(\theta, \alpha) \equiv (\cos \theta (4 + \cos \alpha), -\sin \theta (4 + \cos \alpha), \sin \alpha)^T.$$

Find the area of $\mathbf{f}([0, 2\pi] \times [0, 2\pi])$. **Hint:** Check whether $\mathbf{f}_\theta \cdot \mathbf{f}_\alpha = 0$. This might make the computations reasonable.

15. For $(\theta, \alpha) \in [0, 2\pi] \times [0, 2\pi]$, let

$$\mathbf{f}(\theta, \alpha) \equiv (\cos \theta (3 + 2 \cos \alpha), -\sin \theta (3 + 2 \cos \alpha), 2 \sin \alpha)^T, \quad h(\mathbf{x}) = \cos \alpha,$$

where α is such that $\mathbf{x} = (\cos \theta (3 + 2 \cos \alpha), -\sin \theta (3 + 2 \cos \alpha), 2 \sin \alpha)^T$. Find $\int_{\mathbf{f}([0, 2\pi] \times [0, 2\pi])} h dA$. **Hint:** Check whether $\mathbf{f}_\theta \cdot \mathbf{f}_\alpha = 0$. This might make the computations reasonable.

16. For $(\theta, \alpha) \in [0, 2\pi] \times [0, 2\pi]$, let

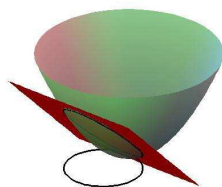
$$\mathbf{f}(\theta, \alpha) \equiv (\cos \theta (4 + 3 \cos \alpha), -\sin \theta (4 + 3 \cos \alpha), 3 \sin \alpha)^T, \quad h(\mathbf{x}) = \cos^2 \theta,$$

where the parametrical description of the surface is

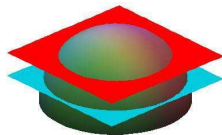
$$\mathbf{x} = (\cos \theta (4 + 3 \cos \alpha), -\sin \theta (4 + 3 \cos \alpha), 3 \sin \alpha)^T$$

Find $\int_{\mathbf{f}([0, 2\pi] \times [0, 2\pi])} h dA$. **Hint:** Check whether $\mathbf{f}_\theta \cdot \mathbf{f}_\alpha = 0$. This might make the computations reasonable.

17. In spherical coordinates, $\phi = c, \rho \in [0, R]$ determines a cone. Find the area of this cone.
18. Let $\mathbf{F} = (x, y, z)$ and let S be the curved surface which comes from the intersection of the plane $z = x$ with the paraboloid $z = x^2 + y^2$. Find an iterated integral for the flux integral $\int_S \mathbf{F} \cdot \mathbf{n} dS$ where \mathbf{n} is the field of unit normals which has negative z component.



19. Let $\mathbf{F} = (x, 0, 0)$ and let S denote the surface which consists of the part of the sphere $x^2 + y^2 + z^2 = 9$ which lies between the planes $z = 1$ and $z = 2$. Find $\int_S \mathbf{F} \cdot \mathbf{n} dS$ where \mathbf{n} is the unit normal to this surface which has positive z component.



20. In the situation of the above problem change the vector field to $\mathbf{F} = (0, 0, z)$ and do the same problem.
21. Show that for a sphere of radius a parameterized with spherical coordinates so that

$$x = a \sin \phi \cos \theta, \quad y = a \sin \phi \sin \theta, \quad z = a \cos \phi$$

the increment of surface area is $a^2 \sin \phi d\theta d\phi$. Use to show that the area of a sphere of radius a is $4\pi a^2$.

Chapter 29

Calculus of Vector Fields

29.1 Divergence and Curl of a Vector Field

Here the important concepts of divergence and curl are defined in terms of rectangular coordinates.

Definition 29.1.1 Let $\mathbf{f} : U \rightarrow \mathbb{R}^p$ for $U \subseteq \mathbb{R}^p$ denote a vector field. A scalar valued function is called a **scalar field**. The function \mathbf{f} is called a C^k **vector field** if the function \mathbf{f} is a C^k function. For a C^1 vector field, as just described $\nabla \cdot \mathbf{f}(\mathbf{x}) \equiv \text{div } \mathbf{f}(\mathbf{x})$ known as the **divergence**, is defined as

$$\nabla \cdot \mathbf{f}(\mathbf{x}) \equiv \text{div } \mathbf{f}(\mathbf{x}) \equiv \sum_{i=1}^p \frac{\partial f_i}{\partial x_i}(\mathbf{x}).$$

Using the repeated summation convention, this is often written as

$$f_{i,i}(\mathbf{x}) \equiv \partial_i f_i(\mathbf{x})$$

where the comma indicates a partial derivative is being taken with respect to the i^{th} variable and ∂_i denotes differentiation with respect to the i^{th} variable. In words, the divergence is the sum of the i^{th} derivative of the i^{th} component function of \mathbf{f} for all values of i . If $p = 3$, the **curl** of the vector field yields another vector field and it is defined as follows.

$$(\text{curl } (\mathbf{f}))(\mathbf{x}) \equiv (\nabla \times \mathbf{f})(\mathbf{x}) \equiv \epsilon_{ijk} \partial_j f_k(\mathbf{x})$$

where here ∂_j means the partial derivative with respect to x_j and the subscript of i in $(\text{curl } (\mathbf{f}))(\mathbf{x})_i$ means the i^{th} Cartesian component of the vector $\text{curl } (\mathbf{f})(\mathbf{x})$. Thus the curl is evaluated by expanding the following determinant along the top row.

$$\begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ f_1(x, y, z) & f_2(x, y, z) & f_3(x, y, z) \end{vmatrix}.$$

Note the similarity with the cross product. Sometimes the curl is called rot. (Short for rotation not decay.) Also

$$\nabla^2 f \equiv \nabla \cdot (\nabla f).$$

This last symbol is important enough that it is given a name, the **Laplacian**. It is also denoted by Δ . Thus $\nabla^2 f = \Delta f$. In addition for \mathbf{f} a vector field, the symbol $\mathbf{f} \cdot \nabla$ is defined as a “differential operator” in the following way.

$$\mathbf{f} \cdot \nabla (g) \equiv f_1(x) \frac{\partial g(x)}{\partial x_1} + f_2(x) \frac{\partial g(x)}{\partial x_2} + \cdots + f_p(x) \frac{\partial g(x)}{\partial x_p}.$$

Thus $\mathbf{f} \cdot \nabla$ takes vector fields and makes them into new vector fields.

This definition is in terms of a given rectangular coordinate system but later coordinate free definitions of the curl and div are presented. For now, everything is defined in terms of a given Cartesian coordinate system. The divergence and curl have profound physical significance and this will be discussed later. For now it is important to understand their definition in terms of coordinates. Be sure you understand that for \mathbf{f} a vector field, $\text{div } \mathbf{f}$ is a scalar field meaning it is a scalar valued function of three variables. For a scalar field f , ∇f is a vector field described earlier. For \mathbf{f} a vector field having values in \mathbb{R}^3 , $\text{curl } \mathbf{f}$ is another vector field.

Example 29.1.2 Let $\mathbf{f}(x) = xy\mathbf{i} + (z-y)\mathbf{j} + (\sin(x)+z)\mathbf{k}$. Find $\text{div } \mathbf{f}$ and $\text{curl } \mathbf{f}$.

First the divergence of \mathbf{f} is

$$\frac{\partial(xy)}{\partial x} + \frac{\partial(z-y)}{\partial y} + \frac{\partial(\sin(x)+z)}{\partial z} = y + (-1) + 1 = y.$$

Now $\text{curl } \mathbf{f}$ is obtained by evaluating

$$\begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ xy & z-y & \sin(x)+z \end{vmatrix} =$$

$$\mathbf{i} \left(\frac{\partial}{\partial y} (\sin(x)+z) - \frac{\partial}{\partial z} (z-y) \right) - \mathbf{j} \left(\frac{\partial}{\partial x} (\sin(x)+z) - \frac{\partial}{\partial z} (xy) \right) +$$

$$\mathbf{k} \left(\frac{\partial}{\partial x} (z-y) - \frac{\partial}{\partial y} (xy) \right) = -\mathbf{i} - \cos(x)\mathbf{j} - x\mathbf{k}.$$

29.1.1 Vector Identities

There are many interesting identities which relate the gradient, divergence and curl.

Theorem 29.1.3 Assuming \mathbf{f}, \mathbf{g} are a C^2 vector fields whenever necessary, the following identities are valid.

1. $\nabla \cdot (\nabla \times \mathbf{f}) = 0$
2. $\nabla \times \nabla \phi = \mathbf{0}$
3. $\nabla \times (\nabla \times \mathbf{f}) = \nabla(\nabla \cdot \mathbf{f}) - \nabla^2 \mathbf{f}$ where $\nabla^2 \mathbf{f}$ is a vector field whose i^{th} component is $\nabla^2 f_i$.
4. $\nabla \cdot (\mathbf{f} \times \mathbf{g}) = \mathbf{g} \cdot (\nabla \times \mathbf{f}) - \mathbf{f} \cdot (\nabla \times \mathbf{g})$

$$5. \nabla \times (\mathbf{f} \times \mathbf{g}) = (\nabla \cdot \mathbf{g}) \mathbf{f} - (\nabla \cdot \mathbf{f}) \mathbf{g} + (\mathbf{g} \cdot \nabla) \mathbf{f} - (\mathbf{f} \cdot \nabla) \mathbf{g}$$

Proof: These are all easy to establish if you use the repeated index summation convention and the reduction identities.

$$\begin{aligned} \nabla \cdot (\nabla \times \mathbf{f}) &= \partial_i (\nabla \times \mathbf{f})_i = \partial_i (\varepsilon_{ijk} \partial_j f_k) = \varepsilon_{ijk} \partial_i (\partial_j f_k) \\ &= \varepsilon_{jik} \partial_j (\partial_i f_k) = -\varepsilon_{ijk} \partial_j (\partial_i f_k) = -\varepsilon_{ijk} \partial_i (\partial_j f_k) \\ &= -\nabla \cdot (\nabla \times \mathbf{f}). \end{aligned}$$

This establishes the first formula. The second formula is done similarly. Now consider the third.

$$\begin{aligned} (\nabla \times (\nabla \times \mathbf{f}))_i &= \varepsilon_{ijk} \partial_j (\nabla \times \mathbf{f})_k = \varepsilon_{ijk} \partial_j (\varepsilon_{krs} \partial_r f_s) \\ &= \overbrace{\varepsilon_{kij}}^{\varepsilon_{ijk}} \varepsilon_{krs} \partial_j (\partial_r f_s) = (\delta_{ir} \delta_{js} - \delta_{is} \delta_{jr}) \partial_j (\partial_r f_s) \\ &= \partial_j (\partial_i f_j) - \partial_j (\partial_j f_i) = \partial_i (\partial_j f_j) - \partial_j (\partial_j f_i) \\ &= (\nabla (\nabla \cdot \mathbf{f}) - \nabla^2 \mathbf{f})_i \end{aligned}$$

This establishes the third identity.

Consider the fourth identity.

$$\begin{aligned} \nabla \cdot (\mathbf{f} \times \mathbf{g}) &= \partial_i (\mathbf{f} \times \mathbf{g})_i = \partial_i \varepsilon_{ijk} f_j g_k \\ &= \varepsilon_{ijk} (\partial_i f_j) g_k + \varepsilon_{ijk} f_j (\partial_i g_k) \\ &= (\varepsilon_{kij} \partial_i f_j) g_k - (\varepsilon_{jik} \partial_i g_k) f_k \\ &= \nabla \times \mathbf{f} \cdot \mathbf{g} - \nabla \times \mathbf{g} \cdot \mathbf{f}. \end{aligned}$$

This proves the fourth identity.

Consider the fifth.

$$\begin{aligned} (\nabla \times (\mathbf{f} \times \mathbf{g}))_i &= \varepsilon_{ijk} \partial_j (\mathbf{f} \times \mathbf{g})_k = \varepsilon_{ijk} \partial_j \varepsilon_{krs} f_r g_s \\ &= \varepsilon_{kij} \varepsilon_{krs} \partial_j (f_r g_s) = (\delta_{ir} \delta_{js} - \delta_{is} \delta_{jr}) \partial_j (f_r g_s) \\ &= \partial_j (f_i g_j) - \partial_j (f_j g_i) \\ &= (\partial_j g_j) f_i + g_j \partial_j f_i - (\partial_j f_j) g_i - f_j (\partial_j g_i) \\ &= ((\nabla \cdot \mathbf{g}) \mathbf{f} + (\mathbf{g} \cdot \nabla) \mathbf{f}) - (\nabla \cdot \mathbf{f}) \mathbf{g} - (\mathbf{f} \cdot \nabla) \mathbf{g})_i \end{aligned}$$

and this establishes the fifth identity. ■

29.1.2 Vector Potentials

One of the above identities says $\nabla \cdot (\nabla \times \mathbf{f}) = 0$. Suppose now $\nabla \cdot \mathbf{g} = 0$. Does it follow that there exists \mathbf{f} such that $\mathbf{g} = \nabla \times \mathbf{f}$? It turns out that this is usually the case and when such an \mathbf{f} exists, it is called a **vector potential**. Here is one way to do it, assuming everything is defined so the following formulas make sense.

$$\mathbf{f}(x, y, z) = \left(\int_0^z g_2(x, y, t) dt, -\int_0^z g_1(x, y, t) dt + \int_0^x g_3(t, y, 0) dt, 0 \right)^T. \quad (29.1)$$

In verifying this you need to use the following manipulation which will generally hold under reasonable conditions but which has not been carefully shown yet.

$$\frac{\partial}{\partial x} \int_a^b h(x, t) dt = \int_a^b \frac{\partial h}{\partial x}(x, t) dt. \quad (29.2)$$

The above formula seems plausible because the integral is a sort of a sum and the derivative of a sum is the sum of the derivatives. However, this sort of sloppy reasoning will get you into all sorts of trouble. The formula involves the interchange of two limit operations, the integral and the limit of a difference quotient. Such an interchange can only be accomplished through a theorem. The following gives the necessary result.

Lemma 29.1.4 Suppose h and $\frac{\partial h}{\partial x}$ are continuous on the rectangle $R = [c, d] \times [a, b]$. Then 29.2 holds.

Proof: Let Δx be such that $x, x + \Delta x$ are both in $[c, d]$. By Theorem 15.12.4 on Page 335 there exists $\delta > 0$ such that if $|(x, t) - (x_1, t_1)| < \delta$, then

$$\left| \frac{\partial h}{\partial x}(x, t) - \frac{\partial h}{\partial x}(x_1, t_1) \right| < \frac{\varepsilon}{b-a}.$$

Let $|\Delta x| < \delta$. Then

$$\begin{aligned} & \left| \int_a^b \frac{h(x + \Delta x, t) - h(x, t)}{\Delta x} dt - \int_a^b \frac{\partial h}{\partial x}(x, t) dt \right| \\ & \leq \int_a^b \left| \frac{h(x + \Delta x, t) - h(x, t)}{\Delta x} - \frac{\partial h}{\partial x}(x, t) \right| dt \\ & = \int_a^b \left| \frac{\partial h(x + \theta_t \Delta x)}{\partial x} - \frac{\partial h}{\partial x}(x, t) \right| dt < \int_a^b \frac{\varepsilon}{b-a} dt = \varepsilon. \end{aligned}$$

Here θ_t is a number between 0 and 1 and going from the second to the third line is an application of the mean value theorem. ■

The second formula of Theorem 29.1.3 states $\nabla \times \nabla \phi = \mathbf{0}$. This suggests the following question: Suppose $\nabla \times \mathbf{f} = \mathbf{0}$, does it follow there exists ϕ , a scalar field such that $\nabla \phi = \mathbf{f}$? The answer to this is often yes and a theorem will be given and proved after the presentation of Stokes' theorem. This scalar field ϕ , is called a **scalar potential** for \mathbf{f} .

29.1.3 The Weak Maximum Principle

There is also a fundamental result having great significance which involves ∇^2 called the maximum principle. This principle says that if $\nabla^2 u \geq 0$ on a bounded open set U , then u achieves its maximum value on the boundary of U .

Theorem 29.1.5 Let U be a bounded open set in \mathbb{R}^p and suppose

$$u \in C^2(U) \cap C(\bar{U})$$

such that $\nabla^2 u \geq 0$ in U . Then letting $\partial U = \bar{U} \setminus U$, it follows that

$$\max \{u(\mathbf{x}) : \mathbf{x} \in \bar{U}\} = \max \{u(\mathbf{x}) : \mathbf{x} \in \partial U\}.$$

Proof: If this is not so, there exists $\mathbf{x}_0 \in U$ such that

$$u(\mathbf{x}_0) > \max \{u(\mathbf{x}) : \mathbf{x} \in \partial U\} \equiv M$$

Since U is bounded, there exists $\varepsilon > 0$ such that

$$u(\mathbf{x}_0) > \max \left\{ u(\mathbf{x}) + \varepsilon |\mathbf{x}|^2 : \mathbf{x} \in \partial U \right\}.$$

Therefore, $u(\mathbf{x}) + \varepsilon |\mathbf{x}|^2$ also has its maximum in U because for ε small enough,

$$u(\mathbf{x}_0) + \varepsilon |\mathbf{x}_0|^2 > u(\mathbf{x}_0) > \max \left\{ u(\mathbf{x}) + \varepsilon |\mathbf{x}|^2 : \mathbf{x} \in \partial U \right\}$$

for all $\mathbf{x} \in \partial U$.

Now let \mathbf{x}_1 be the point in U at which $u(\mathbf{x}) + \varepsilon |\mathbf{x}|^2$ achieves its maximum. As an exercise you should show that $\nabla^2(f+g) = \nabla^2 f + \nabla^2 g$ and therefore, $\nabla^2(u(\mathbf{x}) + \varepsilon |\mathbf{x}|^2) = \nabla^2 u(\mathbf{x}) + 2p\varepsilon$. (Why?) Therefore,

$$0 \geq \nabla^2 u(\mathbf{x}_1) + 2p\varepsilon \geq 2p\varepsilon,$$

a contradiction. ■

29.2 Exercises

1. Find $\operatorname{div} \mathbf{f}$ and $\operatorname{curl} \mathbf{f}$ where \mathbf{f} is

(a) $(xyz, x^2 + \ln(xy), \sin x^2 + z)^T$

(b) $(\sin x, \sin y, \sin z)^T$

(c) $(f(x), g(y), h(z))^T$

(d) $(x-2, y-3, z-6)^T$

(e) $(y^2, 2xy, \cos z)^T$

(f) $(f(y, z), g(x, z), h(y, z))^T$

2. Prove formula 2 of Theorem 29.1.3.

3. Show that if u and v are C^2 functions, then $\operatorname{curl}(u\nabla v) = \nabla u \times \nabla v$.

4. Simplify the expression $\mathbf{f} \times (\nabla \times \mathbf{g}) + \mathbf{g} \times (\nabla \times \mathbf{f}) + (\mathbf{f} \cdot \nabla) \mathbf{g} + (\mathbf{g} \cdot \nabla) \mathbf{f}$.

5. Simplify $\nabla \times (\mathbf{v} \times \mathbf{r})$ where $\mathbf{r} = (x, y, z)^T = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$ and \mathbf{v} is a constant vector.

6. Discover a formula which simplifies $\nabla \cdot (\mathbf{v} \nabla u)$.

7. Verify that $\nabla \cdot (u \nabla v) - \nabla \cdot (v \nabla u) = u \nabla^2 v - v \nabla^2 u$.

8. Verify that $\nabla^2(uv) = v \nabla^2 u + 2(\nabla u \cdot \nabla v) + u \nabla^2 v$.

9. Functions u , which satisfy $\nabla^2 u = 0$ are called harmonic functions. Show that the following functions are harmonic where ever they are defined.

- (a) $2xy$
- (b) $x^2 - y^2$
- (c) $\sin x \cosh y$
- (d) $\ln(x^2 + y^2)$
- (e) $1/\sqrt{x^2 + y^2 + z^2}$

10. Verify the formula given in 29.1 is a vector potential for \mathbf{g} assuming that $\operatorname{div} \mathbf{g} = 0$.

11. Show that if $\nabla^2 u_k = 0$ for each $k = 1, 2, \dots, m$, and c_k is a constant, then

$$\nabla^2 \left(\sum_{k=1}^m c_k u_k \right) = 0$$

also.

12. In Theorem 29.1.5, why is $\nabla^2 (\varepsilon |x|^2) = 2n\varepsilon$?

13. Using Theorem 29.1.5, prove the following: Let $f \in C(\partial U)$ (f is continuous on ∂U .) where U is a bounded open set. Then there exists at most one solution $u \in C^2(U) \cap C(\overline{U})$ and $\nabla^2 u = 0$ in U with $u = f$ on ∂U . **Hint:** Suppose there are two solutions u_i , $i = 1, 2$ and let $w = u_1 - u_2$. Then use the maximum principle.

14. Suppose \mathbf{B} is a vector field and $\nabla \times \mathbf{A} = \mathbf{B}$. Thus \mathbf{A} is a vector potential for \mathbf{B} . Show that $\mathbf{A} + \nabla \phi$ is also a vector potential for \mathbf{B} . Here ϕ is just a C^2 scalar field. Thus the vector potential is not unique.

29.3 The Divergence Theorem

The divergence theorem relates an integral over a set to one on the boundary of the set. It is also called Gauss's theorem.

Definition 29.3.1 A subset V of \mathbb{R}^3 is called cylindrical in the x direction if it is of the form

$$V = \{(x, y, z) : \phi(y, z) \leq x \leq \psi(y, z) \text{ for } (y, z) \in D\}$$

where D is a subset of the yz plane. V is cylindrical in the z direction if

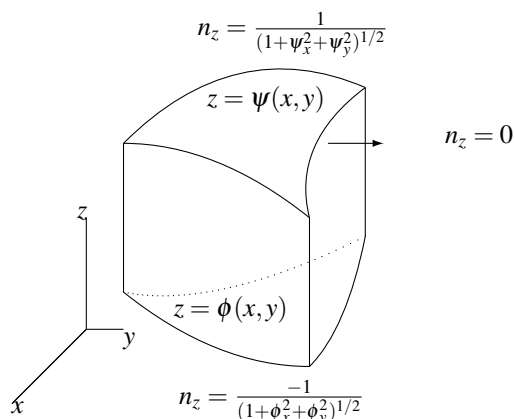
$$V = \{(x, y, z) : \phi(x, y) \leq z \leq \psi(x, y) \text{ for } (x, y) \in D\}$$

where D is a subset of the xy plane, and V is cylindrical in the y direction if

$$V = \{(x, y, z) : \phi(x, z) \leq y \leq \psi(x, z) \text{ for } (x, z) \in D\}$$

where D is a subset of the xz plane. If V is cylindrical in the z direction, denote by ∂V the boundary of V defined as the points of the form $(x, y, \phi(x, y)), (x, y, \psi(x, y))$ for $(x, y) \in D$, along with points of the form (x, y, z) where $(x, y) \in \partial D$ and $\phi(x, y) \leq z \leq \psi(x, y)$. Points on ∂D are defined to be those for which every open ball contains points which are in D as well as points which are not in D . A similar definition holds for ∂V in the case that V is cylindrical in one of the other directions.

The following picture illustrates the above definition in the case of V cylindrical in the z direction. Also labeled are the z components of the respective outer unit normals on the sides and top and bottom.



Of course, many three dimensional sets are cylindrical in each of the coordinate directions. For example, a ball or a rectangle or a tetrahedron are all cylindrical in each direction. The following lemma allows the exchange of the volume integral of a partial derivative for an area integral in which the derivative is replaced with multiplication by an appropriate component of the unit exterior normal.

Lemma 29.3.2 *Suppose V is cylindrical in the z direction and that ϕ and ψ are the functions in the above definition. Assume ϕ and ψ are C^1 functions and suppose F is a C^1 function defined on V . Also, let $\mathbf{n} = (n_x, n_y, n_z)$ be the unit exterior normal to ∂V . Then*

$$\int_V \frac{\partial F}{\partial z}(x, y, z) dV = \int_{\partial V} F n_z dA.$$

Proof: From the fundamental theorem of calculus,

$$\begin{aligned} \int_V \frac{\partial F}{\partial z}(x, y, z) dV &= \int_D \int_{\phi(x, y)}^{\psi(x, y)} \frac{\partial F}{\partial z}(x, y, z) dz dx dy \\ &= \int_D [F(x, y, \psi(x, y)) - F(x, y, \phi(x, y))] dx dy \end{aligned} \quad (29.3)$$

Now the unit exterior normal on the top of V , the surface $(x, y, \psi(x, y))$ is

$$\frac{1}{\sqrt{\psi_x^2 + \psi_y^2 + 1}} (-\psi_x, -\psi_y, 1).$$

This follows from the observation that the top surface is the level surface $z - \psi(x, y) = 0$ and so the gradient of this function of three variables is perpendicular to the level surface. It points in the correct direction because the z component is positive. Therefore, on the top surface

$$n_z = \frac{1}{\sqrt{\psi_x^2 + \psi_y^2 + 1}}$$

Similarly, the unit normal to the surface on the bottom is

$$\frac{1}{\sqrt{\phi_x^2 + \phi_y^2 + 1}} (\phi_x, \phi_y, -1)$$

and so on the bottom surface,

$$n_z = \frac{-1}{\sqrt{\phi_x^2 + \phi_y^2 + 1}}$$

Note that here the z component is negative because since it is the outer normal it must point down. On the lateral surface, the one where $(x, y) \in \partial D$ and $z \in [\phi(x, y), \psi(x, y)]$, $n_z = 0$.

The area element on the top surface is $dA = \sqrt{\psi_x^2 + \psi_y^2 + 1} dx dy$ while the area element on the bottom surface is $\sqrt{\phi_x^2 + \phi_y^2 + 1} dx dy$. Therefore, the last expression in (29.3) is of the form,

$$\begin{aligned} & \int_D F(x, y, \psi(x, y)) \overbrace{\frac{1}{\sqrt{\psi_x^2 + \psi_y^2 + 1}}}^{n_z} \overbrace{\sqrt{\psi_x^2 + \psi_y^2 + 1} dx dy}^{dA} + \\ & \int_D F(x, y, \phi(x, y)) \left(\overbrace{\frac{-1}{\sqrt{\phi_x^2 + \phi_y^2 + 1}}}^{n_z} \right) \overbrace{\sqrt{\phi_x^2 + \phi_y^2 + 1} dx dy}^{dA} \\ & + \int_{\text{Lateral surface}} F n_z dA, \end{aligned}$$

the last term equaling zero because on the lateral surface, $n_z = 0$. Therefore, this reduces to $\int_{\partial V} F n_z dA$ as claimed. ■

The following corollary is entirely similar to the above.

Corollary 29.3.3 *If V is cylindrical in the y direction, then*

$$\int_V \frac{\partial F}{\partial y} dV = \int_{\partial V} F n_y dA$$

and if V is cylindrical in the x direction, then

$$\int_V \frac{\partial F}{\partial x} dV = \int_{\partial V} F n_x dA$$

With this corollary, here is a proof of the divergence theorem.

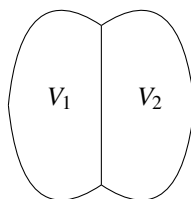
Theorem 29.3.4 *Let V be cylindrical in each of the coordinate directions and let F be a C^1 vector field defined on V . Then*

$$\int_V \nabla \cdot \mathbf{F} dV = \int_{\partial V} \mathbf{F} \cdot \mathbf{n} dA.$$

Proof: From the above lemma and corollary,

$$\begin{aligned}
 \int_V \nabla \cdot \mathbf{F} dV &= \int_V \frac{\partial F_1}{\partial x} + \frac{\partial F_2}{\partial y} + \frac{\partial F_3}{\partial z} dV \\
 &= \int_{\partial V} (F_1 n_x + F_2 n_y + F_3 n_z) dA \\
 &= \int_{\partial V} \mathbf{F} \cdot \mathbf{n} dA. \blacksquare
 \end{aligned}$$

Note that this only requires that ∂V be piecewise continuous. As discussed earlier, the edges end up not contributing to the surface integral. The divergence theorem holds for much more general regions than this. Suppose for example you have a complicated region which is the union of finitely many disjoint regions of the sort just described which are cylindrical in each of the coordinate directions. Then the volume integral over the union of these would equal the sum of the integrals over the disjoint regions. If the boundaries of two of these regions intersect, then the area integrals will cancel out on the intersection because the unit exterior normals will point in opposite directions. Therefore, the sum of the integrals over the boundaries of these disjoint regions will reduce to an integral over the boundary of the union of these. Hence the divergence theorem will continue to hold. For example, consider the following picture. If the divergence theorem holds for each V_i in the following picture, then it holds for the union of these two.



General formulations of the divergence theorem involve Hausdorff measures and the Lebesgue integral, a better integral than the old fashioned Riemann integral which has been obsolete now for almost 100 years. In general, one finds that the conclusion of the divergence theorem is usually true and the theorem can be used with confidence. Minor modifications show that the divergence theorem holds in any dimension. In particular, it holds in two dimensions. In two dimensions, the dS refers to length and the dV refers to area $dx dy$. In four dimensions, the dS would refer to three dimensional area using $dS = \sqrt{1 + \psi_{x_1}^2 + \psi_{x_2}^2 + \psi_{x_3}^2} dx_1 dx_2 dx_3$ or something involving another subset of the four variables on the ends.

Example 29.3.5 Let $V = [0, 1] \times [0, 1] \times [0, 1]$. That is, V is the cube in the first octant having the lower left corner at $(0, 0, 0)$ and the sides of length 1. Let $\mathbf{F}(x, y, z) = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$. Find the flux integral in which \mathbf{n} is the unit exterior normal.

$$\int_{\partial V} \mathbf{F} \cdot \mathbf{n} dS$$

You can certainly inflict much suffering on yourself by breaking the surface up into 6 pieces corresponding to the 6 sides of the cube, finding a parametrization for each face and adding up the appropriate flux integrals. For example, $\mathbf{n} = \mathbf{k}$ on the top face and $\mathbf{n} = -\mathbf{k}$

on the bottom face. On the top face, a parametrization is $(x, y, 1) : (x, y) \in [0, 1] \times [0, 1]$. The area element is just $dx dy$. It is not really all that hard to do it this way but it is much easier to use the divergence theorem. The above integral equals

$$\int_V \operatorname{div}(\mathbf{F}) dV = \int_V 3 dV = 3.$$

Example 29.3.6 This time, let V be the unit ball, $\{(x, y, z) : x^2 + y^2 + z^2 \leq 1\}$ and let

$$\mathbf{F}(x, y, z) = x^2 \mathbf{i} + y \mathbf{j} + (z - 1) \mathbf{k}.$$

Find

$$\int_{\partial V} \mathbf{F} \cdot \mathbf{n} dS.$$

As in the above you could do this by brute force. A parametrization of the ∂V is obtained as

$$x = \sin \phi \cos \theta, \quad y = \sin \phi \sin \theta, \quad z = \cos \phi$$

where $(\phi, \theta) \in (0, \pi) \times (0, 2\pi]$. Now this does not include all the ball but it includes all but the point at the top and at the bottom. As far as the flux integral is concerned these points contribute nothing to the integral so you can neglect them. Then you can grind away and get the flux integral which is desired. However, it is so much easier to use the divergence theorem! Using spherical coordinates,

$$\begin{aligned} \int_{\partial V} \mathbf{F} \cdot \mathbf{n} dS &= \int_V \operatorname{div}(\mathbf{F}) dV = \int_V (2x + 1 + 1) dV \\ &= \int_0^\pi \int_0^{2\pi} \int_0^1 (2 + 2\rho \sin(\phi) \cos \theta) \rho^2 \sin(\phi) d\rho d\theta d\phi = \frac{8}{3} \pi \end{aligned}$$

Example 29.3.7 Suppose V is an open set in \mathbb{R}^3 for which the divergence theorem holds. Let $\mathbf{F}(x, y, z) = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$. Then show that

$$\int_{\partial V} \mathbf{F} \cdot \mathbf{n} dS = 3 \times \text{volume}(V).$$

This follows from the divergence theorem.

$$\int_{\partial V} \mathbf{F} \cdot \mathbf{n} dS = \int_V \operatorname{div}(\mathbf{F}) dV = 3 \int_V dV = 3 \times \text{volume}(V).$$

The message of the divergence theorem is the relation between the volume integral and an area integral. This is the exciting thing about this marvelous theorem. It is not its utility as a method for evaluations of boring problems. This will be shown in the examples of its use which follow.

29.3.1 Coordinate Free Concept of Divergence

The divergence theorem also makes possible a coordinate free definition of the divergence.

Theorem 29.3.8 Let $B(\mathbf{x}, \delta)$ be the ball centered at \mathbf{x} having radius δ and let \mathbf{F} be a C^1 vector field. Then letting $v(B(\mathbf{x}, \delta))$ denote the volume of $B(\mathbf{x}, \delta)$ given by $\int_{B(\mathbf{x}, \delta)} dV$, it follows

$$\operatorname{div} \mathbf{F}(\mathbf{x}) = \lim_{\delta \rightarrow 0^+} \frac{1}{v(B(\mathbf{x}, \delta))} \int_{\partial B(\mathbf{x}, \delta)} \mathbf{F} \cdot \mathbf{n} dA. \quad (29.4)$$

Proof: The divergence theorem holds for balls because they are cylindrical in every direction. Therefore,

$$\frac{1}{v(B(\mathbf{x}, \delta))} \int_{\partial B(\mathbf{x}, \delta)} \mathbf{F} \cdot \mathbf{n} dA = \frac{1}{v(B(\mathbf{x}, \delta))} \int_{B(\mathbf{x}, \delta)} \operatorname{div} \mathbf{F}(\mathbf{y}) dV.$$

Therefore, since $\operatorname{div} \mathbf{F}(\mathbf{x})$ is a constant,

$$\begin{aligned} & \left| \operatorname{div} \mathbf{F}(\mathbf{x}) - \frac{1}{v(B(\mathbf{x}, \delta))} \int_{\partial B(\mathbf{x}, \delta)} \mathbf{F} \cdot \mathbf{n} dA \right| \\ &= \left| \operatorname{div} \mathbf{F}(\mathbf{x}) - \frac{1}{v(B(\mathbf{x}, \delta))} \int_{B(\mathbf{x}, \delta)} \operatorname{div} \mathbf{F}(\mathbf{y}) dV \right| \\ &= \left| \frac{1}{v(B(\mathbf{x}, \delta))} \int_{B(\mathbf{x}, \delta)} (\operatorname{div} \mathbf{F}(\mathbf{x}) - \operatorname{div} \mathbf{F}(\mathbf{y})) dV \right| \\ &\leq \frac{1}{v(B(\mathbf{x}, \delta))} \int_{B(\mathbf{x}, \delta)} |\operatorname{div} \mathbf{F}(\mathbf{x}) - \operatorname{div} \mathbf{F}(\mathbf{y})| dV \\ &\leq \frac{1}{v(B(\mathbf{x}, \delta))} \int_{B(\mathbf{x}, \delta)} \frac{\varepsilon}{2} dV < \varepsilon \end{aligned}$$

whenever ε is small enough, due to the continuity of $\operatorname{div} \mathbf{F}$. Since ε is arbitrary, this shows 29.4. ■

How is this definition independent of coordinates? It only involves geometrical notions of volume and dot product. This is why. Imagine rotating the coordinate axes, keeping all distances the same and expressing everything in terms of the new coordinates. The divergence would still have the same value because of this theorem.

29.4 Applications of the Divergence Theorem

29.4.1 Hydrostatic Pressure

Imagine a fluid which does not move which is acted on by an acceleration \mathbf{g} . Of course the acceleration is usually the acceleration of gravity. Also let the density of the fluid be ρ , a function of position. What can be said about the pressure p in the fluid? Let $B(\mathbf{x}, \varepsilon)$ be a small ball centered at the point \mathbf{x} . Then the force the fluid exerts on this ball would equal

$$-\int_{\partial B(\mathbf{x}, \varepsilon)} p \mathbf{n} dA.$$

Here \mathbf{n} is the unit exterior normal at a small piece of $\partial B(\mathbf{x}, \varepsilon)$ having area dA . By the divergence theorem, (see Problem 1 on Page 612) this integral equals

$$-\int_{B(\mathbf{x}, \varepsilon)} \nabla p dV.$$

Also the force acting on this small ball of fluid is

$$\int_{B(\mathbf{x}, \varepsilon)} \rho \mathbf{g} dV.$$

Since it is given that the fluid does not move, the sum of these forces must equal zero. Thus

$$\int_{B(\mathbf{x}, \varepsilon)} \rho \mathbf{g} dV = \int_{B(\mathbf{x}, \varepsilon)} \nabla p dV.$$

Since this must hold for any ball in the fluid of any radius, it must be that

$$\nabla p = \rho \mathbf{g}. \quad (29.5)$$

It turns out that the pressure in a lake at depth z is equal to $62.5z$. This is easy to see from 29.5. In this case, $\mathbf{g} = g\mathbf{k}$ where $g = 32$ feet/sec². The weight of a cubic foot of water is 62.5 pounds. Therefore, the mass in slugs of this water is $62.5/32$. Since it is a cubic foot, this is also the density of the water in slugs per cubic foot. Also, it is normally assumed that water is incompressible¹. Therefore, this is the mass of water at any depth. Therefore,

$$\frac{\partial p}{\partial x} \mathbf{i} + \frac{\partial p}{\partial y} \mathbf{j} + \frac{\partial p}{\partial z} \mathbf{k} = \frac{62.5}{32} \times 32\mathbf{k}.$$

and so p does not depend on x and y and is only a function of z . It follows $p(0) = 0$, and $p'(z) = 62.5$. Therefore, $p(x, y, z) = 62.5z$. This establishes the claim. This is interesting but 29.5 is more interesting because it does not require ρ to be constant.

29.4.2 Archimedes Law of Buoyancy

Archimedes principle states that when a solid body is immersed in a fluid, the net force acting on the body by the fluid is directly up and equals the total weight of the fluid displaced.

Denote the set of points in three dimensions occupied by the body as V . Then for dA an increment of area on the surface of this body, the force acting on this increment of area would equal $-p dA \mathbf{n}$ where \mathbf{n} is the exterior unit normal. Therefore, since the fluid does not move,

$$\int_{\partial V} -p \mathbf{n} dA = \int_V -\nabla p dV = \int_V \rho g dV \mathbf{k}$$

Which equals the total weight of the displaced fluid and you note the force is directed upward as claimed. Here ρ is the density and 29.5 is being used. There is an interesting point in the above explanation. Why does the second equation hold? Imagine that V were filled with fluid. Then the equation follows from 29.5 because in this equation $\mathbf{g} = -g\mathbf{k}$.

29.4.3 Equations of Heat and Diffusion

Let \mathbf{x} be a point in three dimensional space and let (x_1, x_2, x_3) be Cartesian coordinates of this point. Let there be a three dimensional body having density $\rho = \rho(\mathbf{x}, t)$.

The heat flux \mathbf{J} , in the body is defined as a vector which has the following property.

$$\text{Rate at which heat crosses } S = \int_S \mathbf{J} \cdot \mathbf{n} dA$$

where \mathbf{n} is the unit normal in the desired direction. Thus if V is a three dimensional body,

$$\text{Rate at which heat leaves } V = \int_{\partial V} \mathbf{J} \cdot \mathbf{n} dA$$

¹There is no such thing as an incompressible fluid but this doesn't stop people from making this assumption.

where \mathbf{n} is the unit exterior normal.

Fourier's law of heat conduction states that the heat flux \mathbf{J} satisfies $\mathbf{J} = -k\nabla(u)$ where u is the temperature and $k = k(u, \mathbf{x}, t)$ is called the coefficient of thermal conductivity. This changes depending on the material. It also can be shown by experiment to change with temperature. This equation for the heat flux states that the heat flows from hot places toward colder places in the direction of greatest rate of decrease in temperature. Let $c(\mathbf{x}, t)$ denote the specific heat of the material in the body. This means the amount of heat within V is given by the formula $\int_V \rho(\mathbf{x}, t) c(\mathbf{x}, t) u(\mathbf{x}, t) dV$. Suppose also there are sources for the heat within the material given by $f(\mathbf{x}, u, t)$. If f is positive, the heat is increasing while if f is negative the heat is decreasing. For example such sources could result from a chemical reaction taking place. Then the divergence theorem can be used to verify the following equation for u . Such an equation is called a reaction diffusion equation.

$$\frac{\partial}{\partial t} (\rho(\mathbf{x}, t) c(\mathbf{x}, t) u(\mathbf{x}, t)) = \nabla \cdot (k(u, \mathbf{x}, t) \nabla u(\mathbf{x}, t)) + f(\mathbf{x}, u, t). \quad (29.6)$$

Take an arbitrary V for which the divergence theorem holds. Then the time rate of change of the heat in V is

$$\frac{d}{dt} \int_V \rho(\mathbf{x}, t) c(\mathbf{x}, t) u(\mathbf{x}, t) dV = \int_V \frac{\partial (\rho(\mathbf{x}, t) c(\mathbf{x}, t) u(\mathbf{x}, t))}{\partial t} dV$$

where, as in the preceding example, this is a physical derivation so the consideration of hard mathematics is not necessary. Therefore, from the Fourier law of heat conduction, $\frac{d}{dt} \int_V \rho(\mathbf{x}, t) c(\mathbf{x}, t) u(\mathbf{x}, t) dV =$

$$\begin{aligned} \int_V \frac{\partial (\rho(\mathbf{x}, t) c(\mathbf{x}, t) u(\mathbf{x}, t))}{\partial t} dV &= \overbrace{\int_{\partial V} -\mathbf{J} \cdot \mathbf{n} dA}^{\text{rate at which heat enters}} + \int_V f(\mathbf{x}, u, t) dV \\ &= \int_{\partial V} k \nabla(u) \cdot \mathbf{n} dA + \int_V f(\mathbf{x}, u, t) dV = \int_V (\nabla \cdot (k \nabla(u)) + f) dV. \end{aligned}$$

Since this holds for every sample volume V it must be the case that the above reaction diffusion equation 29.6 holds. Note that more interesting equations can be obtained by letting more of the quantities in the equation depend on temperature. However, the above is a fairly hard equation and people usually assume the coefficient of thermal conductivity depends only on \mathbf{x} and that the reaction term f depends only on \mathbf{x} and t and that ρ and c are constant. Then it reduces to the much easier equation

$$\frac{\partial}{\partial t} u(\mathbf{x}, t) = \frac{1}{\rho c} \nabla \cdot (k(\mathbf{x}) \nabla u(\mathbf{x}, t)) + f(\mathbf{x}, t). \quad (29.7)$$

This is often referred to as the heat equation. Sometimes there are modifications of this in which k is not just a scalar but a matrix to account for different heat flow properties in different directions. However, they are not much harder than the above. The major mathematical difficulties result from allowing k to depend on temperature.

It is known that the heat equation is not correct even if the thermal conductivity did not depend on u because it implies infinite speed of propagation of heat. However, this does not prevent people from using it.

29.4.4 Balance of Mass

Let \mathbf{y} be a point in three dimensional space and let (y_1, y_2, y_3) be Cartesian coordinates of this point. Let V be a region in three dimensional space and suppose a fluid having density $\rho(\mathbf{y}, t)$ and velocity, $\mathbf{v}(\mathbf{y}, t)$ is flowing through this region. Then the mass of fluid leaving V per unit time is given by the area integral $\int_{\partial V} \rho(\mathbf{y}, t) \mathbf{v}(\mathbf{y}, t) \cdot \mathbf{n} dA$ while the total mass of the fluid enclosed in V at a given time is $\int_V \rho(\mathbf{y}, t) dV$. Also suppose mass originates at the rate $f(\mathbf{y}, t)$ per cubic unit per unit time within this fluid. Then the conclusion which can be drawn through the use of the divergence theorem is the following fundamental equation known as the mass balance equation.

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = f(\mathbf{y}, t) \quad (29.8)$$

To see this is so, take an arbitrary V for which the divergence theorem holds. Then the time rate of change of the mass in V is

$$\frac{\partial}{\partial t} \int_V \rho(\mathbf{y}, t) dV = \int_V \frac{\partial \rho(\mathbf{y}, t)}{\partial t} dV$$

where the derivative was taken under the integral sign with respect to t . (This is a physical derivation and therefore, it is not necessary to fuss with the hard mathematics related to the change of limit operations. You should expect this to be true under fairly general conditions because the integral is a sort of sum and the derivative of a sum is the sum of the derivatives.) Therefore, the rate of change of mass $\frac{\partial}{\partial t} \int_V \rho(\mathbf{y}, t) dV$, equals

$$\begin{aligned} \int_V \frac{\partial \rho(\mathbf{y}, t)}{\partial t} dV &= \overbrace{- \int_{\partial V} \rho(\mathbf{y}, t) \mathbf{v}(\mathbf{y}, t) \cdot \mathbf{n} dA}^{\text{rate at which mass enters}} + \int_V f(\mathbf{y}, t) dV \\ &= - \int_V (\nabla \cdot (\rho(\mathbf{y}, t) \mathbf{v}(\mathbf{y}, t)) + f(\mathbf{y}, t)) dV. \end{aligned}$$

Since this holds for every sample volume V it must be the case that the equation of continuity holds. Again, there are interesting mathematical questions here which can be explored but since it is a physical derivation, it is not necessary to dwell too much on them. If all the functions involved are continuous, it is certainly true but it is true under far more general conditions than that.

Also note this equation applies to many situations and f might depend on more than just \mathbf{y} and t . In particular, f might depend also on temperature and the density ρ . This would be the case for example if you were considering the mass of some chemical and f represented a chemical reaction. Mass balance is a general sort of equation valid in many contexts.

29.4.5 Balance of Momentum

This example is a little more substantial than the above. It concerns the balance of momentum for a continuum. To see a full description of all the physics involved, you should consult a book on continuum mechanics. The situation is of a material in three dimensions and it deforms and moves about in three dimensions. This means this material is not a rigid body. Let B_0 denote an open set identifying a chunk of this material at time $t = 0$ and let B_t be an open set which identifies the same chunk of material at time $t > 0$.

Let $\mathbf{y}(t, \mathbf{x}) = (y_1(t, \mathbf{x}), y_2(t, \mathbf{x}), y_3(t, \mathbf{x}))$ denote the position with respect to Cartesian coordinates at time t of the point whose position at time $t = 0$ is $\mathbf{x} = (x_1, x_2, x_3)$. The coordinates \mathbf{x} are sometimes called the reference coordinates and sometimes the material coordinates and sometimes the Lagrangian coordinates. The coordinates \mathbf{y} are called the Eulerian coordinates or sometimes the spacial coordinates and the function $(t, \mathbf{x}) \rightarrow \mathbf{y}(t, \mathbf{x})$ is called the motion. Thus

$$\mathbf{y}(0, \mathbf{x}) = \mathbf{x}. \quad (29.9)$$

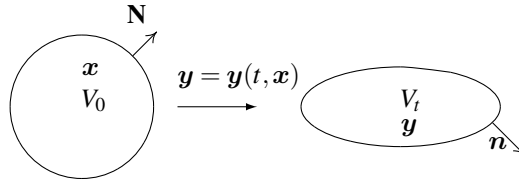
The derivative,

$$D_2 \mathbf{y}(t, \mathbf{x}) \equiv D_{\mathbf{x}} \mathbf{y}(t, \mathbf{x})$$

is called the deformation gradient. Recall the notation means you fix t and consider the function $\mathbf{x} \rightarrow \mathbf{y}(t, \mathbf{x})$, taking its derivative. Since it is a linear transformation, it is represented by the usual matrix, whose i^{th} entry is given by

$$F_{ij}(\mathbf{x}) = \frac{\partial y_i(t, \mathbf{x})}{\partial x_j}.$$

Let $\rho(t, \mathbf{y})$ denote the density of the material at time t at the point \mathbf{y} and let $\rho_0(\mathbf{x})$ denote the density of the material at the point \mathbf{x} . Thus $\rho_0(\mathbf{x}) = \rho(0, \mathbf{x}) = \rho(0, \mathbf{y}(0, \mathbf{x}))$. The first task is to consider the relationship between $\rho(t, \mathbf{y})$ and $\rho_0(\mathbf{x})$. The following picture is useful to illustrate the ideas.



Lemma 29.4.1 $\rho_0(\mathbf{x}) = \rho(t, \mathbf{y}(t, \mathbf{x})) \det(F)$ and in any reasonable physical motion $\det(F) > 0$.

Proof: Let V_0 represent a small chunk of material at $t = 0$ and let V_t represent the same chunk of material at time t . I will be a little sloppy and refer to V_0 as the small chunk of material at time $t = 0$ and V_t as the chunk of material at time t rather than an open set representing the chunk of material. Then by the change of variables formula for multiple integrals,

$$\int_{V_t} dV = \int_{V_0} |\det(F)| dV.$$

If $\det(F) = 0$ for some t the above formula shows that the chunk of material went from positive volume to zero volume and this is not physically possible. Therefore, it is impossible that $\det(F)$ can equal zero. However, at $t = 0$, $F = I$, the identity because of 29.9. Therefore, $\det(F) = 1$ at $t = 0$ and if it is assumed $t \rightarrow \det(F)$ is continuous it follows by the intermediate value theorem that $\det(F) > 0$ for all t . ■

Of course it is not known for sure that this function is continuous but the above shows why it is at least reasonable to expect $\det(F) > 0$. General arguments involve measure and the Lebesgue integral. As usual, one neglects mathematical rigor in derivations of physical formulae.

Now using the change of variables formula

$$\begin{aligned}\text{mass of } V_t &= \int_{V_t} \rho(t, \mathbf{y}) dV(\mathbf{y}) = \int_{V_0} \rho(t, \mathbf{y}(t, \mathbf{x})) \det(F) dV(\mathbf{x}) \\ &= \text{mass of } V_0 = \int_{V_0} \rho_0(\mathbf{x}) dV.\end{aligned}$$

Since V_0 is arbitrary, it follows

$$\rho_0(\mathbf{x}) = \rho(t, \mathbf{y}(t, \mathbf{x})) \det(F)$$

as claimed. Note this shows that $\det(F)$ is a magnification factor for the density.

Now consider a small chunk of material, V_t at time t which corresponds to V_0 at time $t = 0$. The total linear momentum of this material at time t is

$$\int_{V_t} \rho(t, \mathbf{y}) \mathbf{v}(t, \mathbf{y}) dV$$

where \mathbf{v} is the velocity. By Newton's second law, the time rate of change of this linear momentum should equal the total force acting on the chunk of material. In the following derivation, $dV(\mathbf{y})$ will indicate the integration is taking place with respect to the variable \mathbf{y} . By Lemma 29.4.1 and the change of variables formula for multiple integrals

$$\begin{aligned}& \frac{d}{dt} \left(\int_{V_t} \rho(t, \mathbf{y}) \mathbf{v}(t, \mathbf{y}) dV(\mathbf{y}) \right) \\ &= \frac{d}{dt} \left(\int_{V_0} \rho(t, \mathbf{y}(t, \mathbf{x})) \mathbf{v}(t, \mathbf{y}(t, \mathbf{x})) \det(F) dV(\mathbf{x}) \right) \\ &= \frac{d}{dt} \left(\int_{V_0} \rho_0(\mathbf{x}) \mathbf{v}(t, \mathbf{y}(t, \mathbf{x})) dV(\mathbf{x}) \right) = \int_{V_0} \rho_0(\mathbf{x}) \left[\frac{\partial \mathbf{v}}{\partial t} + \frac{\partial \mathbf{v}}{\partial y_i} \frac{\partial y_i}{\partial t} \right] dV(\mathbf{x}) \\ &= \int_{V_0} \rho_0(\mathbf{x}) \frac{1}{\det(F)} \left[\frac{\partial \mathbf{v}}{\partial t} + \frac{\partial \mathbf{v}}{\partial y_i} \frac{\partial y_i}{\partial t} \right] \det(F) dV(\mathbf{x}) \\ &= \int_{V_0} \overbrace{\rho(t, \mathbf{y}(t, \mathbf{x})) \det(F)}^{=\rho_0(\mathbf{x})} \frac{1}{\det(F)} \left[\frac{\partial \mathbf{v}}{\partial t} + \frac{\partial \mathbf{v}}{\partial y_i} \frac{\partial y_i}{\partial t} \right] \det(F) dV(\mathbf{y}) \\ &= \int_{V_0} \rho(t, \mathbf{y}(t, \mathbf{x})) \left[\frac{\partial \mathbf{v}}{\partial t} + \frac{\partial \mathbf{v}}{\partial y_i} \frac{\partial y_i}{\partial t} \right] \det(F) dV(\mathbf{y}) \\ &= \int_{V_t} \rho(t, \mathbf{y}) \left[\frac{\partial \mathbf{v}}{\partial t} + \frac{\partial \mathbf{v}}{\partial y_i} \frac{\partial y_i}{\partial t} \right] dV(\mathbf{y}) = \int_{V_t} \rho(t, \mathbf{y}) \dot{\mathbf{v}} dV(\mathbf{y})\end{aligned}$$

where the dot on \mathbf{v} indicates it is the total derivative. Having taken the derivative of the total momentum, it is time to consider the total force acting on the chunk of material.

The force comes from two sources, a body force \mathbf{b} and a force which acts on the boundary of the chunk of material called a traction force. Typically, the body force is something like gravity in which case, $\mathbf{b} = -g\rho\mathbf{k}$, assuming the Cartesian coordinate system has been chosen in the usual manner. The traction force is of the form

$$\int_{\partial V_t} \mathbf{s}(t, \mathbf{y}, \mathbf{n}) dA$$

where \mathbf{n} is the unit exterior normal. Thus the traction force depends on position, time, and the orientation of the boundary of V_t . Cauchy showed the existence of a linear transformation $T(t, \mathbf{y})$ such that $T(t, \mathbf{y})\mathbf{n} = \mathbf{s}(t, \mathbf{y}, \mathbf{n})$. It follows there is a matrix $T_{ij}(t, \mathbf{y})$ such that the i^{th} component of \mathbf{s} is given by $s_i(t, \mathbf{y}, \mathbf{n}) = T_{ij}(t, \mathbf{y})n_j$. Cauchy also showed this matrix is symmetric, $T_{ij} = T_{ji}$. It is called the Cauchy stress. Using Newton's second law to equate the time derivative of the total linear momentum with the applied forces and using the usual repeated index summation convention,

$$\int_{V_t} \rho(t, \mathbf{y}) \dot{\mathbf{v}} dV(\mathbf{y}) = \int_{V_t} \mathbf{b}(t, \mathbf{y}) dV(\mathbf{y}) + \int_{\partial B_t} \mathbf{e}_i T_{ij}(t, \mathbf{y}) n_j dA,$$

the sum taken over repeated indices. Here is where the divergence theorem is used. In the last integral, the multiplication by n_j is exchanged for the j^{th} partial derivative and an integral over V_t . Thus

$$\int_{V_t} \rho(t, \mathbf{y}) \dot{\mathbf{v}} dV(\mathbf{y}) = \int_{V_t} \mathbf{b}(t, \mathbf{y}) dV(\mathbf{y}) + \int_{V_t} \frac{\mathbf{e}_i \partial (T_{ij}(t, \mathbf{y}))}{\partial y_j} dV(\mathbf{y}),$$

the sum taken over repeated indices. Since V_t was arbitrary, it follows

$$\rho(t, \mathbf{y}) \dot{\mathbf{v}} = \mathbf{b}(t, \mathbf{y}) + \mathbf{e}_i \frac{\partial (T_{ij}(t, \mathbf{y}))}{\partial y_j} \equiv \mathbf{b}(t, \mathbf{y}) + \text{div}(\mathbf{T})$$

where here $\text{div} \mathbf{T}$ is a vector whose i^{th} component is given by

$$(\text{div} \mathbf{T})_i = \frac{\partial T_{ij}}{\partial y_j}.$$

The term $\frac{\partial \mathbf{v}}{\partial t} + \frac{\partial \mathbf{v}}{\partial y_i} \frac{\partial y_i}{\partial t}$, is the total derivative with respect to t of the velocity \mathbf{v} . Thus you might see this written as

$$\rho \dot{\mathbf{v}} = \mathbf{b} + \text{div}(\mathbf{T}).$$

The above formulation of the balance of momentum involves the spatial coordinates \mathbf{y} but people also like to formulate momentum balance in terms of the material coordinates \mathbf{x} . Of course this changes everything.

The momentum in terms of the material coordinates is

$$\int_{V_0} \rho_0(\mathbf{x}) \mathbf{v}(t, \mathbf{x}) dV$$

and so, since \mathbf{x} does not depend on t ,

$$\frac{d}{dt} \left(\int_{V_0} \rho_0(\mathbf{x}) \mathbf{v}(t, \mathbf{x}) dV \right) = \int_{V_0} \rho_0(\mathbf{x}) \mathbf{v}_t(t, \mathbf{x}) dV.$$

As indicated earlier, this is a physical derivation, so the mathematical questions related to interchange of limit operations are ignored. This must equal the total applied force. Thus using the repeated index summation convention,

$$\int_{V_0} \rho_0(\mathbf{x}) \mathbf{v}_t(t, \mathbf{x}) dV = \int_{V_0} \mathbf{b}_0(t, \mathbf{x}) dV + \int_{\partial V_t} \mathbf{e}_i T_{ij} n_j dA, \quad (29.10)$$

the first term on the right being the contribution of the body force given per unit volume in the material coordinates and the last term being the traction force discussed earlier. The task is to write this last integral as one over ∂V_0 . For $\mathbf{y} \in \partial V_t$ there is a unit outer normal \mathbf{n} . Here $\mathbf{y} = \mathbf{y}(t, \mathbf{x})$ for $\mathbf{x} \in \partial V_0$. Then define \mathbf{N} to be the unit outer normal to V_0 at the point \mathbf{x} . Near the point $\mathbf{y} \in \partial V_t$ the surface ∂V_t is given parametrically in the form $\mathbf{y} = \mathbf{y}(s, t)$ for $(s, t) \in D \subseteq \mathbb{R}^2$ and it can be assumed the unit normal to ∂V_t near this point is

$$\mathbf{n} = \frac{\mathbf{y}_s(s, t) \times \mathbf{y}_t(s, t)}{|\mathbf{y}_s(s, t) \times \mathbf{y}_t(s, t)|}$$

with the area element given by $|\mathbf{y}_s(s, t) \times \mathbf{y}_t(s, t)| ds dt$. This is true for $\mathbf{y} \in P_t \subseteq \partial V_t$, a small piece of ∂V_t . Therefore, the last integral in 29.10 is the sum of integrals over small pieces of the form

$$\int_{P_t} T_{ij} n_j dA \quad (29.11)$$

where P_t is parameterized by $\mathbf{y}(s, t)$, $(s, t) \in D$. Thus the integral in 29.11 is of the form

$$\int_D T_{ij}(\mathbf{y}(s, t)) (\mathbf{y}_s(s, t) \times \mathbf{y}_t(s, t))_j ds dt.$$

By the chain rule this equals

$$\int_D T_{ij}(\mathbf{y}(s, t)) \left(\frac{\partial \mathbf{y}}{\partial x_\alpha} \frac{\partial x_\alpha}{\partial s} \times \frac{\partial \mathbf{y}}{\partial x_\beta} \frac{\partial x_\beta}{\partial t} \right)_j ds dt.$$

Summation over repeated indices is used. Remember $\mathbf{y} = \mathbf{y}(t, \mathbf{x})$ and it is always assumed the mapping $\mathbf{x} \rightarrow \mathbf{y}(t, \mathbf{x})$ is one to one and so, since on the surface ∂V_t near \mathbf{y} , the points are functions of (s, t) , it follows \mathbf{x} is also a function of (s, t) . Now by the properties of the cross product, this last integral equals

$$\int_D T_{ij}(\mathbf{x}(s, t)) \frac{\partial x_\alpha}{\partial s} \frac{\partial x_\beta}{\partial t} \left(\frac{\partial \mathbf{y}}{\partial x_\alpha} \times \frac{\partial \mathbf{y}}{\partial x_\beta} \right)_j ds dt \quad (29.12)$$

where here $\mathbf{x}(s, t)$ is the point of ∂V_0 which corresponds with $\mathbf{y}(s, t) \in \partial V_t$. Thus

$$T_{ij}(\mathbf{x}(s, t)) = T_{ij}(\mathbf{y}(s, t)).$$

(Perhaps this is a slight abuse of notation because T_{ij} is defined on ∂V_t , not on ∂V_0 , but it avoids introducing extra symbols.) Next 29.12 equals

$$\begin{aligned} & \int_D T_{ij}(\mathbf{x}(s, t)) \frac{\partial x_\alpha}{\partial s} \frac{\partial x_\beta}{\partial t} \varepsilon_{jab} \frac{\partial y_a}{\partial x_\alpha} \frac{\partial y_b}{\partial x_\beta} ds dt \\ &= \int_D T_{ij}(\mathbf{x}(s, t)) \frac{\partial x_\alpha}{\partial s} \frac{\partial x_\beta}{\partial t} \varepsilon_{cab} \delta_{jc} \frac{\partial y_a}{\partial x_\alpha} \frac{\partial y_b}{\partial x_\beta} ds dt \\ &= \int_D T_{ij}(\mathbf{x}(s, t)) \frac{\partial x_\alpha}{\partial s} \frac{\partial x_\beta}{\partial t} \varepsilon_{cab} \overbrace{\frac{\partial y_c}{\partial x_p} \frac{\partial x_p}{\partial y_j}}^{=\delta_{jc}} \frac{\partial y_a}{\partial x_\alpha} \frac{\partial y_b}{\partial x_\beta} ds dt \end{aligned}$$

$$\begin{aligned}
&= \int_D T_{ij}(\mathbf{x}(s, t)) \frac{\partial x_\alpha}{\partial s} \frac{\partial x_\beta}{\partial t} \frac{\partial x_p}{\partial y_j} \overbrace{\epsilon_{cab} \frac{\partial y_c}{\partial x_p} \frac{\partial y_a}{\partial x_\alpha} \frac{\partial y_b}{\partial x_\beta}}^{=\epsilon_{p\alpha\beta} \det(F)} ds dt \\
&= \int_D (\det F) T_{ij}(\mathbf{x}(s, t)) \epsilon_{p\alpha\beta} \frac{\partial x_\alpha}{\partial s} \frac{\partial x_\beta}{\partial t} \frac{\partial x_p}{\partial y_j} ds dt.
\end{aligned}$$

Now $\frac{\partial x_p}{\partial y_j} = F_{pj}^{-1}$ and also

$$\epsilon_{p\alpha\beta} \frac{\partial x_\alpha}{\partial s} \frac{\partial x_\beta}{\partial t} = (\mathbf{x}_s \times \mathbf{x}_t)_p$$

so the result just obtained is of the form

$$\begin{aligned}
&\int_D (\det F) F_{pj}^{-1} T_{ij}(\mathbf{x}(s, t)) (\mathbf{x}_s \times \mathbf{x}_t)_p ds dt = \\
&\int_D (\det F) T_{ij}(\mathbf{x}(s, t)) (F^{-T})_{jp} (\mathbf{x}_s \times \mathbf{x}_t)_p ds dt.
\end{aligned}$$

This has transformed the integral over P_t to one over P_0 , the part of ∂V_0 which corresponds with P_t . Thus the last integral is of the form

$$\int_{P_0} \det(F) (TF^{-T})_{ip} N_p dA$$

Summing these up over the pieces of ∂V_t and ∂V_0 , yields the last integral in 29.10 equals

$$\int_{\partial V_0} \det(F) (TF^{-T})_{ip} N_p dA$$

and so the balance of momentum in terms of the material coordinates becomes

$$\int_{V_0} \rho_0(\mathbf{x}) \mathbf{v}_t(t, \mathbf{x}) dV = \int_{V_0} \mathbf{b}_0(t, \mathbf{x}) dV + \int_{\partial V_0} \mathbf{e}_i \det(F) (TF^{-T})_{ip} N_p dA$$

The matrix $\det(F) (TF^{-T})_{ip}$ is called the Piola Kirchhoff stress S . An application of the divergence theorem yields

$$\int_{V_0} \rho_0(\mathbf{x}) \mathbf{v}_t(t, \mathbf{x}) dV = \int_{V_0} \mathbf{b}_0(t, \mathbf{x}) dV + \int_{V_0} \mathbf{e}_i \frac{\partial (\det(F) (TF^{-T})_{ip})}{\partial x_p} dV.$$

Since V_0 is arbitrary, a balance law for momentum in terms of the material coordinates is obtained

$$\begin{aligned}
\rho_0(\mathbf{x}) \mathbf{v}_t(t, \mathbf{x}) &= \mathbf{b}_0(t, \mathbf{x}) + \mathbf{e}_i \frac{\partial (\det(F) (TF^{-T})_{ip})}{\partial x_p} \\
&= \mathbf{b}_0(t, \mathbf{x}) + \operatorname{div} (\det(F) (TF^{-T})) \\
&= \mathbf{b}_0(t, \mathbf{x}) + \operatorname{div} S.
\end{aligned} \tag{29.13}$$

As just shown, the relation between the Cauchy stress and the Piola Kirchhoff stress is

$$S = \det(F) (TF^{-T}), \tag{29.14}$$

perhaps not the first thing you would think of.

The main purpose of this presentation is to show how the divergence theorem is used in a significant way to obtain balance laws and to indicate a very interesting direction for further study. To continue, one needs to specify T or S as an appropriate function of things related to the motion \mathbf{y} . Often the thing related to the motion is something called the strain and such relationships are known as constitutive laws.

29.4.6 Frame Indifference

The proper formulation of constitutive laws involves more physical considerations such as frame indifference in which it is required that the response of the system cannot depend on the manner in which the Cartesian coordinate system for the spacial coordinates was chosen.

For $Q(t)$ an orthogonal transformation,

$$\mathbf{y}' = \mathbf{q}(t) + Q(t) \mathbf{y},$$

the new spacial coordinates are denoted by \mathbf{y}' . Recall an orthogonal transformation is just one which satisfies

$$Q(t)^T Q(t) = Q(t) Q(t)^T = I.$$

The stress has to do with the traction force area density produced by internal changes in the body and has nothing to do with the way the body is observed. Therefore, it is required that

$$T' \mathbf{n}' = Q T \mathbf{n}$$

Thus

$$T' Q \mathbf{n} = Q T \mathbf{n}$$

Since this is true for any \mathbf{n} normal to the boundary of any piece of the material considered, it must be the case that

$$T' Q = Q T$$

and so

$$T' = Q T Q^T.$$

This is called frame indifference.

By 29.14, the Piola Kirchhoff stress S is related to T by

$$S = \det(F) T F^{-T}, \quad F \equiv D_{\mathbf{x}} \mathbf{y}.$$

This stress involves the use of the material coordinates and a normal \mathbf{N} to a piece of the body in reference configuration. Thus $S \mathbf{N}$ gives the force on a part of ∂V_t per unit area on ∂V_0 . Then for a different choice of spacial coordinates, $\mathbf{y}' = \mathbf{q}(t) + Q(t) \mathbf{y}$,

$$S' = \det(F') T' (F')^{-T}$$

but

$$F' = D_{\mathbf{x}} \mathbf{y}' = Q(t) D_{\mathbf{x}} \mathbf{y} = Q F$$

and so frame indifference in terms of S is

$$S' = \det(F) Q T Q^T (Q F)^{-T} = \det(F) Q T Q^T Q F^{-T} = Q S$$

This principle of frame indifference is sometimes ignored and there are certainly interesting mathematical models which have resulted from doing this, but such things cannot be considered physically acceptable.

There are also many other physical properties which can be included, which require a certain form for the constitutive equations. These considerations are outside the scope of this book and require a considerable amount of linear algebra.

There are also balance laws for energy which you may study later but these are more problematic than the balance laws for mass and momentum. However, the divergence theorem is used in these also.

29.4.7 Bernoulli's Principle

Consider a possibly moving fluid with constant density ρ and let P denote the pressure in this fluid. If B is a part of this fluid the force exerted on B by the rest of the fluid is $\int_{\partial B} -P \mathbf{n} dA$ where \mathbf{n} is the outer normal from B . Assume this is the only force which matters so for example there is no viscosity in the fluid. Thus the Cauchy stress in rectangular coordinates should be

$$T = \begin{pmatrix} -P & 0 & 0 \\ 0 & -P & 0 \\ 0 & 0 & -P \end{pmatrix}.$$

Then $\operatorname{div} T = -\nabla P$. Also suppose the only body force is from gravity, a force of the form $-\rho g \mathbf{k}$, so from the balance of momentum

$$\rho \dot{\mathbf{v}} = -\rho g \mathbf{k} - \nabla P(\mathbf{x}). \quad (29.15)$$

Now in all this, the coordinates are the spacial coordinates, and it is assumed they are rectangular. Thus $\mathbf{x} = (x, y, z)^T$ and \mathbf{v} is the velocity while $\dot{\mathbf{v}}$ is the total derivative of $\mathbf{v} = (v_1, v_2, v_3)^T$ given by $\mathbf{v}_t + v_i \mathbf{v}_{,i}$. Take the dot product of both sides of (29.15) with \mathbf{v} . This yields

$$(\rho/2) \frac{d}{dt} |\mathbf{v}|^2 = -\rho g \frac{dz}{dt} - \frac{d}{dt} P(\mathbf{x}).$$

Therefore,

$$\frac{d}{dt} \left(\frac{\rho |\mathbf{v}|^2}{2} + \rho g z + P(\mathbf{x}) \right) = 0,$$

so there is a constant C' such that

$$\frac{\rho |\mathbf{v}|^2}{2} + \rho g z + P(\mathbf{x}) = C'$$

For convenience define γ to be the weight density of this fluid. Thus $\gamma = \rho g$. Divide by γ . Then

$$\frac{|\mathbf{v}|^2}{2g} + z + \frac{P(\mathbf{x})}{\gamma} = C.$$

This is Bernoulli's² principle. Note how, if you keep the height the same, then if you raise $|\mathbf{v}|$, it follows the pressure drops.

²There were many Bernoullis. This is Daniel Bernoulli. He seems to have been nicer than some of the others. Daniel was actually a doctor who was interested in mathematics. He lived from 1700-1782.

This is often used to explain the lift of an airplane wing. The top surface is curved, which forces the air to go faster over the top of the wing, causing a drop in pressure which creates lift. It is also used to explain the concept of a venturi tube in which the air loses pressure due to being pinched which causes it to flow faster. In many of these applications, the assumptions used in which ρ is constant, and there is no other contribution to the traction force on ∂B than pressure, so in particular, there is no viscosity, are not correct. However, it is hoped that the effects of these deviations from the ideal situation are small enough that the conclusions are still roughly true. You can see how using balance of momentum can be used to consider more difficult situations. For example, you might have a body force which is more involved than gravity.

29.4.8 The Wave Equation

As an example of how the balance law of momentum is used to obtain an important equation of mathematical physics, suppose $S = kF$ where k is a constant and F is the deformation gradient and let $\mathbf{u} \equiv \mathbf{y} - \mathbf{x}$. Thus \mathbf{u} is the displacement. Then from 29.13 you can verify the following holds.

$$\rho_0(\mathbf{x}) \mathbf{u}_{tt}(t, \mathbf{x}) = \mathbf{b}_0(t, \mathbf{x}) + k\Delta \mathbf{u}(t, \mathbf{x}) \quad (29.16)$$

In the case where ρ_0 is a constant and $\mathbf{b}_0 = 0$, this yields

$$\mathbf{u}_{tt} - c\Delta \mathbf{u} = \mathbf{0}.$$

The wave equation is $u_{tt} - c\Delta u = 0$ and so the above gives three wave equations, one for each component.

29.4.9 A Negative Observation

Many of the above applications of the divergence theorem are based on the assumption that matter is continuously distributed in a way that the above arguments are correct. In other words, a continuum. However, there is no such thing as a continuum. It has been known for some time now that matter is composed of atoms. It is not continuously distributed through some region of space as it is in the above. Apologists for this contradiction with reality sometimes say to consider enough of the material in question that it is reasonable to think of it as a continuum. This mystical reasoning is then violated as soon as they go from the integral form of the balance laws to the differential equations expressing the traditional formulation of these laws. See Problem 10 below, for example. However, these laws continue to be used and seem to lead to useful physical models which have value in predicting the behavior of physical systems. This is what justifies their use, not any fundamental truth.

29.4.10 Volumes of Balls in \mathbb{R}^n

Recall, $B(\mathbf{x}, r)$ denotes the set of all $\mathbf{y} \in \mathbb{R}^n$ such that $|\mathbf{y} - \mathbf{x}| < r$. 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 $\alpha_n r^n$ where α_n will denote the volume of

the unit ball in \mathbb{R}^n . 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. By the divergence theorem,

$$\int_{B(\mathbf{0}, r)} \operatorname{div} \mathbf{x} \, dx = \int_{\partial B(\mathbf{0}, r)} \mathbf{x} \cdot \frac{\mathbf{x}}{|\mathbf{x}|} dA$$

because the unit outward normal on $\partial B(\mathbf{0}, r)$ is $\frac{\mathbf{x}}{|\mathbf{x}|}$. Therefore, denoting $A(\partial B)$ as the area of ∂B ,

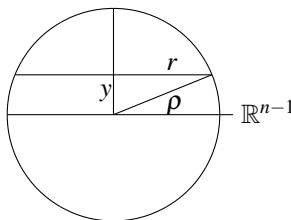
$$n\alpha_n r^n = rA(\partial B(\mathbf{0}, r))$$

and so

$$A(\partial B(\mathbf{0}, r)) = n\alpha_n r^{n-1}.$$

You recall the surface area of $S^2 \equiv \{\mathbf{x} \in \mathbb{R}^3 : |\mathbf{x}| = r\}$ 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_n denote the area of the sphere $S^{n-1} = \{\mathbf{x} \in \mathbb{R}^n : |\mathbf{x}| = 1\}$. I just showed that $\omega_n = n\alpha_n$.

I want to find α_n now and also to get a relationship between ω_n and ω_{n-1} . Consider the following picture of the ball of radius ρ seen on the side.



Taking slices at height y as shown and using that these slices have $n-1$ dimensional area equal to $\alpha_{n-1} r^{n-1}$, it follows

$$\alpha_n \rho^n = 2 \int_0^\rho \alpha_{n-1} (\rho^2 - y^2)^{(n-1)/2} dy$$

In the integral, change variables, letting $y = \rho \cos \theta$. Then

$$\alpha_n \rho^n = 2\rho^n \alpha_{n-1} \int_0^{\pi/2} \sin^n(\theta) d\theta.$$

It follows that

$$\alpha_n = 2\alpha_{n-1} \int_0^{\pi/2} \sin^n(\theta) d\theta. \quad (29.17)$$

Consequently,

$$\omega_n = \frac{2n\omega_{n-1}}{n-1} \int_0^{\pi/2} \sin^n(\theta) d\theta. \quad (29.18)$$

This is a little messier than I would like.

$$\begin{aligned} \int_0^{\pi/2} \sin^n(\theta) d\theta &= -\cos \theta \sin^{n-1} \theta \Big|_0^{\pi/2} + (n-1) \int_0^{\pi/2} \cos^2 \theta \sin^{n-2} \theta \\ &= (n-1) \int_0^{\pi/2} (1 - \sin^2 \theta) \sin^{n-2}(\theta) d\theta \\ &= (n-1) \int_0^{\pi/2} \sin^{n-2}(\theta) d\theta - (n-1) \int_0^{\pi/2} \sin^n(\theta) d\theta \end{aligned}$$

Hence

$$n \int_0^{\pi/2} \sin^n(\theta) d\theta = (n-1) \int_0^{\pi/2} \sin^{n-2}(\theta) d\theta \quad (29.19)$$

and so 29.18 is of the form

$$\omega_n = 2\omega_{n-1} \int_0^{\pi/2} \sin^{n-2}(\theta) d\theta. \quad (29.20)$$

So what is α_n explicitly? Clearly $\alpha_1 = 2$ and $\alpha_2 = \pi$.

Theorem 29.4.2 $\alpha_n = \frac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)}$ where Γ denotes the gamma function, defined for $\alpha > 0$ by

$$\Gamma(\alpha) \equiv \int_0^\infty e^{-t} t^{\alpha-1} dt.$$

Proof: Recall that $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$. Now note the given formula holds if $n = 1$ because

$$\Gamma\left(\frac{1}{2}+1\right) = \frac{1}{2}\Gamma\left(\frac{1}{2}\right) = \frac{\sqrt{\pi}}{2}.$$

I leave it as an exercise for you to verify that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. Thus $\alpha_1 = 2 = \frac{\sqrt{\pi}}{\sqrt{\pi}/2}$ satisfying the formula. Now suppose this formula holds for $k \leq n$. Then from the induction hypothesis, 29.20, 29.19, 29.17 and 29.18,

$$\begin{aligned} \alpha_{n+1} &= 2\alpha_n \int_0^{\pi/2} \sin^{n+1}(\theta) d\theta = 2\alpha_n \frac{n}{n+1} \int_0^{\pi/2} \sin^{n-1}(\theta) d\theta \\ &= 2\alpha_n \frac{n}{n+1} \frac{\alpha_{n-1}}{2\alpha_{n-2}} = \frac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)} \frac{n}{n+1} \pi^{1/2} \frac{\Gamma(\frac{n-2}{2}+1)}{\Gamma(\frac{n-1}{2}+1)} \\ &= \frac{\pi^{n/2}}{\Gamma(\frac{n-2}{2}+1)} \frac{n}{\binom{n}{2}} \frac{n}{n+1} \pi^{1/2} \frac{\Gamma(\frac{n-2}{2}+1)}{\Gamma(\frac{n-1}{2}+1)} \\ &= 2\pi^{(n+1)/2} \frac{1}{n+1} \frac{1}{\Gamma(\frac{n-1}{2}+1)} = \pi^{(n+1)/2} \frac{1}{\binom{n+1}{2}} \frac{1}{\Gamma(\frac{n-1}{2}+1)} \\ &= \pi^{(n+1)/2} \frac{1}{\binom{n+1}{2} \Gamma(\frac{n+1}{2})} = \frac{\pi^{(n+1)/2}}{\Gamma(\frac{n+1}{2}+1)}. \quad \blacksquare \end{aligned}$$

29.4.11 Electrostatics

Coloumb's law says that the electric field intensity at \mathbf{x} of a charge q located at point \mathbf{x}_0 is given by

$$\mathbf{E} = k \frac{q(\mathbf{x} - \mathbf{x}_0)}{|\mathbf{x} - \mathbf{x}_0|^3}$$

where the electric field intensity is defined to be the force experienced by a unit positive charge placed at the point \mathbf{x} . Note that this is a vector and that its direction depends on the sign of q . It points away from \mathbf{x}_0 if q is positive and points toward \mathbf{x}_0 if q is negative. The constant k is a physical constant like the gravitation constant. It has been computed through careful experiments similar to those used with the calculation of the gravitation constant.

The interesting thing about Coloumb's law is that \mathbf{E} is the gradient of a function. In fact,

$$\mathbf{E} = \nabla \left(qk \frac{1}{|\mathbf{x} - \mathbf{x}_0|} \right).$$

The other thing which is significant about this is that in three dimensions and for $\mathbf{x} \neq \mathbf{x}_0$,

$$\nabla \cdot \nabla \left(qk \frac{1}{|\mathbf{x} - \mathbf{x}_0|} \right) = \nabla \cdot \mathbf{E} = 0. \quad (29.21)$$

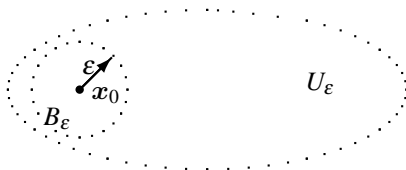
This is left as an exercise for you to verify.

These observations will be used to derive a very important formula for the integral

$$\int_{\partial U} \mathbf{E} \cdot \mathbf{n} dS$$

where \mathbf{E} is the electric field intensity due to a charge, q located at the point $\mathbf{x}_0 \in U$, a bounded open set for which the divergence theorem holds.

Let U_ε denote the open set obtained by removing the open ball centered at \mathbf{x}_0 which has radius ε where ε is small enough that the following picture is a correct representation of the situation.



Then on the boundary of B_ε the unit outer normal to U_ε is $-\frac{\mathbf{x} - \mathbf{x}_0}{|\mathbf{x} - \mathbf{x}_0|}$. Therefore,

$$\begin{aligned} \int_{\partial B_\varepsilon} \mathbf{E} \cdot \mathbf{n} dS &= - \int_{\partial B_\varepsilon} k \frac{q(\mathbf{x} - \mathbf{x}_0)}{|\mathbf{x} - \mathbf{x}_0|^3} \cdot \frac{\mathbf{x} - \mathbf{x}_0}{|\mathbf{x} - \mathbf{x}_0|} dS \\ &= -kq \int_{\partial B_\varepsilon} \frac{1}{|\mathbf{x} - \mathbf{x}_0|^2} dS = \frac{-kq}{\varepsilon^2} \int_{\partial B_\varepsilon} dS \\ &= \frac{-kq}{\varepsilon^2} 4\pi\varepsilon^2 = -4\pi kq. \end{aligned}$$

Therefore, from the divergence theorem and observation 29.21,

$$-4\pi kq + \int_{\partial U} \mathbf{E} \cdot \mathbf{n} dS = \int_{\partial U_\varepsilon} \mathbf{E} \cdot \mathbf{n} dS = \int_{U_\varepsilon} \nabla \cdot \mathbf{E} dV = 0.$$

It follows that $4\pi kq = \int_{\partial U} \mathbf{E} \cdot \mathbf{n} dS$. If there are several charges located inside U , say q_1, q_2, \dots, q_n , then letting \mathbf{E}_i denote the electric field intensity of the i^{th} charge and \mathbf{E} denoting the total resulting electric field intensity due to all these charges,

$$\int_{\partial U} \mathbf{E} \cdot \mathbf{n} dS = \sum_{i=1}^n \int_{\partial U} \mathbf{E}_i \cdot \mathbf{n} dS = \sum_{i=1}^n 4\pi kq_i = 4\pi k \sum_{i=1}^n q_i.$$

This is known as Gauss's law and it is the fundamental result in electrostatics.

29.5 Exercises

1. 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 $\int_{\partial V} \mathbf{n} u dA = \int_V \nabla u dV$. Note this implies $\int_V \frac{\partial u}{\partial x} dV = \int_{\partial V} n_1 u dA$.
2. Fick's law for diffusion states the flux of a diffusing species, \mathbf{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.
3. Sometimes people consider diffusion in materials which are not homogeneous. This means that $\mathbf{J} = -K \nabla c$ where K is a 3×3 matrix. 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.
4. Let V be such that the divergence theorem holds. Show that $\int_V \nabla \cdot (u \nabla v) dV = \int_{\partial V} u \frac{\partial v}{\partial \mathbf{n}} dA$ where \mathbf{n} is the exterior normal and $\frac{\partial v}{\partial \mathbf{n}}$ denotes the directional derivative of v in the direction \mathbf{n} .

5. Let V be such that the divergence theorem holds. Show that

$$\int_V (v \nabla^2 u - u \nabla^2 v) dV = \int_{\partial V} \left(v \frac{\partial u}{\partial \mathbf{n}} - u \frac{\partial v}{\partial \mathbf{n}} \right) dA$$

where \mathbf{n} is the exterior normal and $\frac{\partial u}{\partial \mathbf{n}}$ is defined in Problem 4.

6. 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 4 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 $h(t) = w(t\mathbf{x} + (1-t)\mathbf{y})$ and showing h is a constant. Alternatively, you might consider the maximum principle.
7. Show that $\int_{\partial V} \nabla \times \mathbf{v} \cdot \mathbf{n} dA = 0$ where V is a region for which the divergence theorem holds and \mathbf{v} is a C^2 vector field.
8. Let $\mathbf{F}(x, y, z) = (x, y, z)$ be a vector field in \mathbb{R}^3 and let V be a three dimensional shape and let $\mathbf{n} = (n_1, n_2, n_3)$. Show that $\int_{\partial V} (xn_1 + yn_2 + zn_3) dA = 3 \times \text{volume of } V$.
9. Let $\mathbf{F} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$ 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.
10. Suppose $f : U \rightarrow \mathbb{R}$ is continuous where U is some open set and for all $B \subseteq U$ where B is a ball, $\int_B f(\mathbf{x}) dV = 0$. Show that this implies $f(\mathbf{x}) = 0$ for all $\mathbf{x} \in U$.

11. 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 $\int_{\partial U} \mathbf{F} \cdot \mathbf{n} dS$ where $\mathbf{F} = (x+3, 2y, 3z)$. **Hint:** If you like, you might want to use the divergence theorem.
12. Find the flux out of the cylinder whose base is $x^2 + y^2 \leq 1$ which has height 2 of the vector field $\mathbf{F} = (xy, zy, z^2 + x)$.
13. Find the flux out of the ball of radius 4 centered at $\mathbf{0}$ of the vector field $\mathbf{F} = (x, zy, z+x)$.
14. Verify 29.16 from 29.13 and the assumption that $S = kF$.
15. Show that if $u_k, k = 1, 2, \dots, n$ each satisfies 29.7 with $f = 0$ then for any choice of constants c_1, \dots, c_n , so does $\sum_{k=1}^n c_k u_k$.
16. Suppose $k(\mathbf{x}) = k$, a constant and $f = 0$. Then 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 equation³
17. Let U be a three dimensional region for which the divergence theorem holds. Show that $\int_U \nabla \times \mathbf{F} dx = \int_{\partial U} \mathbf{n} \times \mathbf{F} dS$ where \mathbf{n} is the unit outer normal.
18. In a linear, viscous, incompressible fluid, the Cauchy stress is of the form

$$T_{ij}(t, \mathbf{y}) = \lambda \left(\frac{v_{i,j}(t, \mathbf{y}) + v_{j,i}(t, \mathbf{y})}{2} \right) - p \delta_{ij}$$

where p is the pressure, δ_{ij} equals 0 if $i \neq j$ and 1 if $i = j$, and the comma followed by an index indicates the partial derivative with respect to that variable and \mathbf{v} is the velocity. Thus $v_{i,j} = \frac{\partial v_i}{\partial y_j}$. Also, p denotes the pressure. Show, using the balance of mass equation that incompressible implies $\text{div } \mathbf{v} = 0$. Next show that the balance of momentum equation requires

$$\rho \dot{\mathbf{v}} - \frac{\lambda}{2} \Delta \mathbf{v} = \rho \left[\frac{\partial \mathbf{v}}{\partial t} + \frac{\partial \mathbf{v}}{\partial y_i} v_i \right] - \frac{\lambda}{2} \Delta \mathbf{v} = \mathbf{b} - \nabla p.$$

This is the famous Navier Stokes equation for incompressible viscous linear fluids. There are still open questions related to this equation, one of which is worth \$1,000,000 at this time.

³Fourier, an officer in Napoleon's army studied solutions to the heat equation back in 1813. He was interested in heat flow in cannons. He sought to find solutions by adding up infinitely many multiples of solutions of this form, the multiples coming from a given initial condition occurring when $t = 0$. Fourier thought that the resulting series always converged to this function. Lagrange and Laplace were not so sure. This topic of Fourier series, especially the question of convergence, fascinated mathematicians for the next 150 years and motivated the development of analysis. The first proof of convergence was given by Dirichlet. As mentioned earlier, Dirichlet, Riemann, and later Lebesgue and Fejer were all interested in the convergence of Fourier series and the last big result on convergence did not take place till the middle 1960's and was due to Carleson and more generally by Hunt. It was a surprise because of a negative result of Kolmogorov from 1923. Actually these ideas were used by many others before Fourier, but the idea became associated with him.

Fourier was with Napoleon in Egypt when the Rosetta Stone was discovered and wrote about Egypt in Description de l'Égypte. He was a teacher of Champollion who eventually made it possible to read Egyptian by using the Rosetta Stone. This expedition of Napoleon caused great interest in all things Egyptian in the first part of the nineteenth century.

Chapter 30

Stokes and Green's Theorems

30.1 Green's Theorem

Green's theorem is an important theorem which relates line integrals to integrals over a surface in the plane. It can be used to establish the seemingly more general Stoke's theorem but is interesting for it's own sake. Historically, theorems like it were important in the development of complex analysis.

Here is a proof of Green's theorem from the divergence theorem.

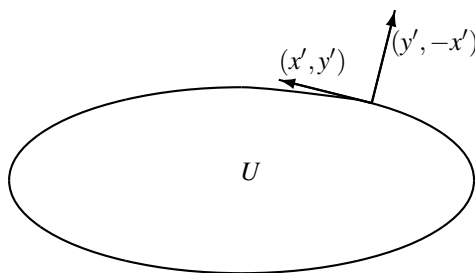
Theorem 30.1.1 (*Green's Theorem*) Let U be an open set in the plane for which the divergence theorem holds. Let ∂U be piecewise smooth, and let

$$\mathbf{F}(x,y) = (P(x,y), Q(x,y))$$

be a C^1 vector field defined near U . Then

$$\int_{\partial U} \mathbf{F} \cdot d\mathbf{R} = \int_U \left(\frac{\partial Q}{\partial x}(x,y) - \frac{\partial P}{\partial y}(x,y) \right) dA.$$

Proof: Suppose the divergence theorem holds for U . Consider the following picture.



Since it is assumed that motion around U is counter clockwise, the tangent vector (x', y') is as shown. The unit **exterior normal** is a multiple of

$$(x', y', 0) \times (0, 0, 1) = (y', -x', 0).$$

Use your right hand and the geometric description of the cross product to verify this. This would be the case at all the points where the unit exterior normal exists.

Now let $\mathbf{F}(x, y) = (Q(x, y), -P(x, y))$. Also note the area (length) element on the bounding curve ∂U is $\sqrt{(x')^2 + (y')^2} dt$. Suppose the boundary of U consists of m smooth curves, the i^{th} of which is parameterized by (x_i, y_i) with the parameter $t \in [a_i, b_i]$. Then by the divergence theorem,

$$\begin{aligned} \int_U (Q_x - P_y) dA &= \int_U \operatorname{div}(\mathbf{F}) dA = \int_{\partial U} \mathbf{F} \cdot \mathbf{n} dS \\ &= \sum_{i=1}^m \int_{a_i}^{b_i} (Q(x_i(t), y_i(t)), -P(x_i(t), y_i(t))) \\ &\quad \cdot \frac{1}{\sqrt{(x'_i)^2 + (y'_i)^2}} \overbrace{(y'_i, -x'_i) \sqrt{(x'_i)^2 + (y'_i)^2}}^{dS} dt \\ &= \sum_{i=1}^m \int_{a_i}^{b_i} (Q(x_i(t), y_i(t)), -P(x_i(t), y_i(t))) \cdot (y'_i, -x'_i) dt \\ &= \sum_{i=1}^m \int_{a_i}^{b_i} Q(x_i(t), y_i(t)) y'_i(t) + P(x_i(t), y_i(t)) x'_i(t) dt \equiv \int_{\partial U} P dx + Q dy \end{aligned}$$

This proves Green's theorem from the divergence theorem. ■

Proposition 30.1.2 *Let U be an open set in \mathbb{R}^2 for which Green's theorem holds. Then Area of $U = \int_{\partial U} \mathbf{F} \cdot d\mathbf{R}$ where $\mathbf{F}(x, y) = \frac{1}{2}(-y, x)$, $(0, x)$, or $(-y, 0)$.*

Proof: This follows immediately from Green's theorem. ■

Example 30.1.3 *Use Proposition 30.1.2 to find the area of the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} \leq 1$.*

You can parameterize the boundary of this ellipse as $x = a \cos t$, $y = b \sin t$, $t \in [0, 2\pi]$. Then from Proposition 30.1.2,

$$\text{Area equals} = \frac{1}{2} \int_0^{2\pi} (-b \sin t, a \cos t) \cdot (-a \sin t, b \cos t) dt = \frac{1}{2} \int_0^{2\pi} (ab) dt = \pi ab.$$

Example 30.1.4 *Find $\int_{\partial U} \mathbf{F} \cdot d\mathbf{R}$ where U is the set $\{(x, y) : x^2 + 3y^2 \leq 9\}$ and $\mathbf{F}(x, y) = (y, -x)$.*

One way to do this is to parameterize the boundary of U and then compute the line integral directly. It is easier to use Green's theorem. The desired line integral equals $\int_U ((-1) - 1) dA = -2 \int_U dA$. Now U is an ellipse having area equal to $3\sqrt{3}$ and so the answer is $-6\sqrt{3}$.

Example 30.1.5 *Find $\int_{\partial U} \mathbf{F} \cdot d\mathbf{R}$ where U is the set $\{(x, y) : 2 \leq x \leq 4, 0 \leq y \leq 3\}$ and*

$$\mathbf{F}(x, y) = (x \sin y, y^3 \cos x)$$

From Green's theorem this line integral equals

$$\int_2^4 \int_0^3 (-y^3 \sin x - x \cos y) dy dx = \frac{81}{4} \cos 4 - 6 \sin 3 - \frac{81}{4} \cos 2.$$

This is much easier than computing the line integral because you don't have to break the boundary in pieces and consider each separately.

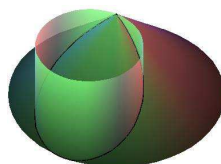
Example 30.1.6 Find $\int_{\partial U} \mathbf{F} \cdot d\mathbf{R}$ where U is the set $\{(x, y) : 2 \leq x \leq 4, x \leq y \leq 4\}$ and

$$\mathbf{F}(x, y) = (x \sin y, y \sin x)$$

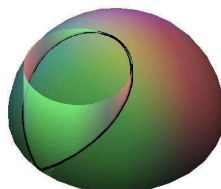
From Green's theorem, this line integral equals $\int_2^4 \int_x^4 (y \cos x - x \cos y) dy dx = 4 \cos 2 - 8 \cos 4 - 8 \sin 2 - 4 \sin 4$.

30.2 Exercises

1. Find $\int_S x dS$ where S is the surface which results from the intersection of the cone $z = 2 - \sqrt{x^2 + y^2}$ with the cylinder $x^2 + y^2 - 2x = 0$.

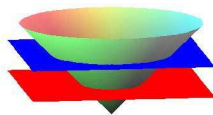


2. Now let \mathbf{n} be the unit normal to the above surface which has positive z component and let $\mathbf{F}(x, y, z) = (x, y, z)$. Find the flux integral $\int_S \mathbf{F} \cdot \mathbf{n} dS$.
3. Find $\int_S z dS$ where S is the surface which results from the intersection of the hemisphere $z = \sqrt{4 - x^2 - y^2}$ with the cylinder $x^2 + y^2 - 2x = 0$.



4. In the situation of the above problem, find the flux integral $\int_S \mathbf{F} \cdot \mathbf{n} dS$ where \mathbf{n} is the unit normal to the surface which has positive z component and $\mathbf{F} = (x, y, z)$.
5. Let $x^2/a^2 + y^2/b^2 = 1$ be an ellipse. Show using Green's theorem that its area is πab .
6. A spherical storage tank having radius a is filled with water which weights 62.5 pounds per cubic foot. It is shown later that this implies that the pressure of the water at depth z equals $62.5z$. Find the total force acting on this storage tank.

7. Let \mathbf{n} be the unit normal to the cone $z = \sqrt{x^2 + y^2}$ which has negative z component and let $\mathbf{F} = (x, 0, z)$ be a vector field. Let S be the part of this cone which lies between the planes $z = 1$ and $z = 2$.

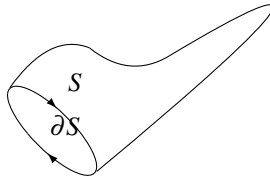


Find $\int_S \mathbf{F} \cdot \mathbf{n} dS$.

8. Let S be the surface $z = 9 - x^2 - y^2$ for $x^2 + y^2 \leq 9$. Let \mathbf{n} be the unit normal to S which points up. Let $\mathbf{F} = (y, -x, z)$ and find $\int_S \mathbf{F} \cdot \mathbf{n} dS$.
9. Let S be the surface $3z = 9 - x^2 - y^2$ for $x^2 + y^2 \leq 9$. Let \mathbf{n} be the unit normal to S which points up. Let $\mathbf{F} = (y, -x, z)$ and find $\int_S \mathbf{F} \cdot \mathbf{n} dS$.
10. For $\mathbf{F} = (x, y, z)$, S is the part of the cylinder $x^2 + y^2 = 1$ between the planes $z = 1$ and $z = 3$. Letting \mathbf{n} be the unit normal which points away from the z axis, find $\int_S \mathbf{F} \cdot \mathbf{n} dS$.
11. Let S be the part of the sphere of radius a which lies between the two cones $\phi = \frac{\pi}{4}$ and $\phi = \frac{\pi}{6}$. Let $\mathbf{F} = (z, y, 0)$. Find the flux integral $\int_S \mathbf{F} \cdot \mathbf{n} dS$.
12. Let S be the part of a sphere of radius a above the plane $z = \frac{a}{2}$, $\mathbf{F} = (2x, 1, 1)$ and let \mathbf{n} be the unit upward normal on S . Find $\int_S \mathbf{F} \cdot \mathbf{n} dS$.
13. In the above, problem, let C be the boundary of S oriented counter clockwise as viewed from high on the z axis. Find $\int_C 2x dx + dy + dz$.
14. Let S be the top half of a sphere of radius a centered at $\mathbf{0}$ and let \mathbf{n} be the unit outward normal. Let $\mathbf{F} = (0, 0, z)$. Find $\int_S \mathbf{F} \cdot \mathbf{n} dS$.
15. Let D be a circle in the plane which has radius 1 and let C be its counter clockwise boundary. Find $\int_C y dx + x dy$.
16. Let D be a circle in the plane which has radius 1 and let C be its counter clockwise boundary. Find $\int_C y dx - x dy$.
17. Find $\int_C (x + y) dx$ where C is the square curve which goes from $(0, 0) \rightarrow (1, 0) \rightarrow (1, 1) \rightarrow (0, 1) \rightarrow (0, 0)$.
18. Find the line integral $\int_C (\sin x + y) dx + y^2 dy$ where C is the oriented square $(0, 0) \rightarrow (1, 0) \rightarrow (1, 1) \rightarrow (0, 1) \rightarrow (0, 0)$.
19. Let $P(x, y) = \frac{-y}{x^2 + y^2}$, $Q(x, y) = \frac{x}{x^2 + y^2}$. Show $Q_x - P_y = 0$. Let D be the unit disk. Compute directly $\int_C P dx + Q dy$ where C is the counter clockwise circle of radius 1 which bounds the unit disk. Why don't you get 0 for the line integral?
20. Let $\mathbf{F} = (2y, \ln(1 + y^2) + x)$. Find $\int_C \mathbf{F} \cdot d\mathbf{R}$ where C is the curve consisting of line segments, $(0, 0) \rightarrow (1, 0) \rightarrow (1, 1) \rightarrow (0, 0)$.

30.3 Stoke's Theorem from Green's Theorem

Stoke's theorem is a generalization of Green's theorem which relates the integral over a surface to the integral around the boundary of the surface. These terms are a little different from what occurs in \mathbb{R}^2 . To describe this, consider a sock. The surface is the sock and its boundary will be the edge of the opening of the sock in which you place your foot. Another way to think of this is to imagine a region in \mathbb{R}^2 of the sort discussed above for Green's theorem. Suppose it is on a sheet of rubber and the sheet of rubber is stretched in three dimensions. The boundary of the resulting surface is the result of the stretching applied to the boundary of the original region in \mathbb{R}^2 . Here is a picture describing the situation.



Recall the following definition of the curl of a vector field. Why do we even consider it?

Definition 30.3.1 Let $\mathbf{F}(x, y, z) = (F_1(x, y, z), F_2(x, y, z), F_3(x, y, z))$ be a C^1 vector field defined on an open set V in \mathbb{R}^3 . Then

$$\nabla \times \mathbf{F} \equiv \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ F_1 & F_2 & F_3 \end{vmatrix} \equiv \left(\frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z} \right) \mathbf{i} + \left(\frac{\partial F_1}{\partial z} - \frac{\partial F_3}{\partial x} \right) \mathbf{j} + \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right) \mathbf{k}.$$

This is also called $\text{curl}(\mathbf{F})$ and written as indicated, $\nabla \times \mathbf{F}$.

The following lemma gives the fundamental identity which will be used in the proof of Stoke's theorem.

Lemma 30.3.2 Let $\mathbf{R}: U \rightarrow V \subseteq \mathbb{R}^3$ where U is an open subset of \mathbb{R}^2 and V is an open subset of \mathbb{R}^3 . Suppose \mathbf{R} is C^2 and let \mathbf{F} be a C^1 vector field defined in V .

$$(\mathbf{R}_u \times \mathbf{R}_v) \cdot (\nabla \times \mathbf{F})(\mathbf{R}(u, v)) = ((\mathbf{F} \circ \mathbf{R})_u \cdot \mathbf{R}_v - (\mathbf{F} \circ \mathbf{R})_v \cdot \mathbf{R}_u)(u, v). \quad (30.1)$$

Proof: Start with the left side and let $x_i = R_i(u, v)$ for short.

$$\begin{aligned} (\mathbf{R}_u \times \mathbf{R}_v) \cdot (\nabla \times \mathbf{F})(\mathbf{R}(u, v)) &= \varepsilon_{ijk} x_{ju} x_{kv} \varepsilon_{irs} \frac{\partial F_s}{\partial x_r} = (\delta_{jr} \delta_{ks} - \delta_{js} \delta_{kr}) x_{ju} x_{kv} \frac{\partial F_s}{\partial x_r} \\ &= x_{ju} x_{kv} \frac{\partial F_k}{\partial x_j} - x_{ju} x_{kv} \frac{\partial F_j}{\partial x_k} = \mathbf{R}_v \cdot \frac{\partial (\mathbf{F} \circ \mathbf{R})}{\partial u} - \mathbf{R}_u \cdot \frac{\partial (\mathbf{F} \circ \mathbf{R})}{\partial v} \end{aligned}$$

which proves 30.1. ■

The proof of Stoke's theorem given next follows [10]. First, it is convenient to give a definition.

Definition 30.3.3 A vector valued function $\mathbf{R}: U \subseteq \mathbb{R}^m \rightarrow \mathbb{R}^n$ is said to be in $C^k(\overline{U}, \mathbb{R}^n)$ if it is the restriction to \overline{U} of a vector valued function which is defined on \mathbb{R}^m and is C^k . That is, this function has continuous partial derivatives up to order k .

Theorem 30.3.4 (Stoke's Theorem) Let U be any region in \mathbb{R}^2 for which the conclusion of Green's theorem holds and let $\mathbf{R} \in C^2(\bar{U}, \mathbb{R}^3)$ be a one to one function satisfying $|(\mathbf{R}_u \times \mathbf{R}_v)(u, v)| \neq 0$ for all $(u, v) \in U$ and let S denote the surface

$$S \equiv \{\mathbf{R}(u, v) : (u, v) \in U\}, \quad \partial S \equiv \{\mathbf{R}(u, v) : (u, v) \in \partial U\}$$

where the orientation on ∂S is consistent with the counter clockwise orientation on ∂U (U is on the left as you walk around ∂U). Then for \mathbf{F} a C^1 vector field defined near S ,

$$\int_{\partial S} \mathbf{F} \cdot d\mathbf{R} = \int_S \text{curl}(\mathbf{F}) \cdot \mathbf{n} dS$$

where \mathbf{n} is the normal to S defined by

$$\mathbf{n} \equiv \frac{\mathbf{R}_u \times \mathbf{R}_v}{|\mathbf{R}_u \times \mathbf{R}_v|}.$$

Proof: Letting C be an oriented part of ∂U having parametrization, $\mathbf{r}(t) \equiv (u(t), v(t))$ for $t \in [\alpha, \beta]$ and letting $\mathbf{R}(C)$ denote the oriented part of ∂S corresponding to C , $\int_{\mathbf{R}(C)} \mathbf{F} \cdot d\mathbf{R} =$

$$\begin{aligned} &= \int_{\alpha}^{\beta} \mathbf{F}(\mathbf{R}(u(t), v(t))) \cdot (\mathbf{R}_u u'(t) + \mathbf{R}_v v'(t)) dt \\ &= \int_{\alpha}^{\beta} \mathbf{F}(\mathbf{R}(u(t), v(t))) \mathbf{R}_u(u(t), v(t)) u'(t) dt \\ &\quad + \int_{\alpha}^{\beta} \mathbf{F}(\mathbf{R}(u(t), v(t))) \mathbf{R}_v(u(t), v(t)) v'(t) dt \\ &= \int_C ((\mathbf{F} \circ \mathbf{R}) \cdot \mathbf{R}_u, (\mathbf{F} \circ \mathbf{R}) \cdot \mathbf{R}_v) \cdot d\mathbf{r}. \end{aligned}$$

Since this holds for each such piece of ∂U , it follows

$$\int_{\partial S} \mathbf{F} \cdot d\mathbf{R} = \int_{\partial U} ((\mathbf{F} \circ \mathbf{R}) \cdot \mathbf{R}_u, (\mathbf{F} \circ \mathbf{R}) \cdot \mathbf{R}_v) \cdot d\mathbf{r}.$$

By the assumption that the conclusion of Green's theorem holds for U , this equals

$$\begin{aligned} &\int_U [((\mathbf{F} \circ \mathbf{R}) \cdot \mathbf{R}_v)_u - ((\mathbf{F} \circ \mathbf{R}) \cdot \mathbf{R}_u)_v] dA \\ &= \int_U [(\mathbf{F} \circ \mathbf{R})_u \cdot \mathbf{R}_v + (\mathbf{F} \circ \mathbf{R}) \cdot \mathbf{R}_{vu} - (\mathbf{F} \circ \mathbf{R}) \cdot \mathbf{R}_{uv} - (\mathbf{F} \circ \mathbf{R})_v \cdot \mathbf{R}_u] dA \\ &= \int_U [(\mathbf{F} \circ \mathbf{R})_u \cdot \mathbf{R}_v - (\mathbf{F} \circ \mathbf{R})_v \cdot \mathbf{R}_u] dA \end{aligned}$$

the last step holding by equality of mixed partial derivatives, a result of the assumption that \mathbf{R} is C^2 . Now by Lemma 30.3.2, this equals

$$\begin{aligned} &\int_U (\mathbf{R}_u \times \mathbf{R}_v) \cdot (\nabla \times \mathbf{F}) dA \\ &= \int_U \nabla \times \mathbf{F} \cdot (\mathbf{R}_u \times \mathbf{R}_v) dA = \int_S \nabla \times \mathbf{F} \cdot \mathbf{n} dS \end{aligned}$$

because $dS = |(\mathbf{R}_u \times \mathbf{R}_v)| dA$ and $\mathbf{n} = \frac{(\mathbf{R}_u \times \mathbf{R}_v)}{|(\mathbf{R}_u \times \mathbf{R}_v)|}$. Thus

$$(\mathbf{R}_u \times \mathbf{R}_v) dA = \frac{(\mathbf{R}_u \times \mathbf{R}_v)}{|(\mathbf{R}_u \times \mathbf{R}_v)|} |(\mathbf{R}_u \times \mathbf{R}_v)| dA = \mathbf{n} dS.$$

This proves Stoke's theorem. ■

Note that there is no mention made in the final result that \mathbf{R} is C^2 . Therefore, it is not surprising that versions of this theorem are valid in which this assumption is not present. It is possible to obtain extremely general versions of Stoke's theorem if you use the Lebesgue integral.

30.3.1 The Normal and the Orientation

Stoke's theorem as just presented needs no apology. However, it is helpful in applications to have some additional geometric insight.

To begin with, suppose the surface S of interest is a parallelogram in \mathbb{R}^3 determined by the two vectors \mathbf{a}, \mathbf{b} . Thus $S = \mathbf{R}(Q)$ where $Q = [0, 1] \times [0, 1]$ is the unit square and for $(u, v) \in Q$,

$$\mathbf{R}(u, v) \equiv u\mathbf{a} + v\mathbf{b} + \mathbf{p},$$

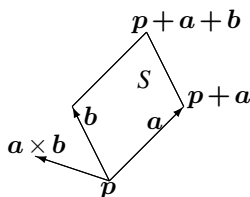
the point \mathbf{p} being a corner of the parallelogram S . Then orient ∂S consistent with the counter clockwise orientation on ∂Q . Thus, following this orientation on S you go from \mathbf{p} to $\mathbf{p} + \mathbf{a}$ to $\mathbf{p} + \mathbf{a} + \mathbf{b}$ to $\mathbf{p} + \mathbf{b}$ to \mathbf{p} . Then Stoke's theorem implies that with this orientation on ∂S ,

$$\int_{\partial S} \mathbf{F} \cdot d\mathbf{R} = \int_S \nabla \times \mathbf{F} \cdot \mathbf{n} ds$$

where

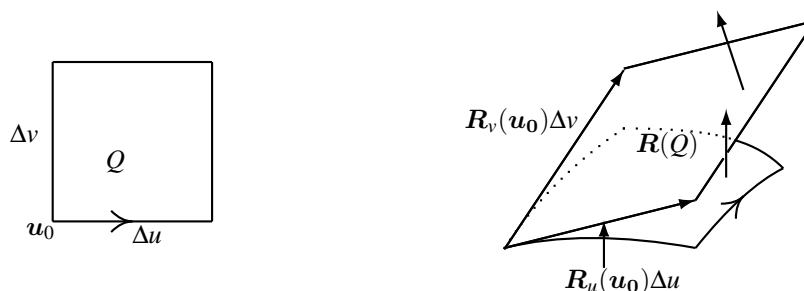
$$\mathbf{n} = \mathbf{R}_u \times \mathbf{R}_v / |\mathbf{R}_u \times \mathbf{R}_v| = \mathbf{a} \times \mathbf{b} / |\mathbf{a} \times \mathbf{b}|.$$

Now recall $\mathbf{a}, \mathbf{b}, \mathbf{a} \times \mathbf{b}$ forms a right hand system.



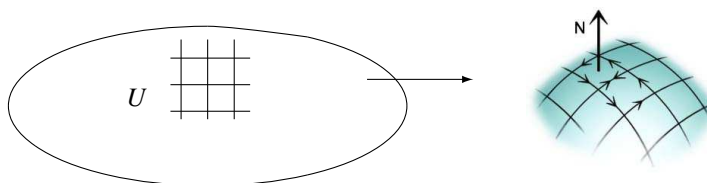
Thus, if you were walking around ∂S in the direction of the orientation with your left hand over the surface S , the normal vector $\mathbf{a} \times \mathbf{b}$ would be pointing in the direction of your head.

More generally, if S is a surface which is not necessarily a parallelogram but is instead as described in Theorem 30.3.4, you could consider a **small** rectangle Q contained in U and orient the boundary of $\mathbf{R}(Q)$ consistent with the counter clockwise orientation on ∂Q . Then if Q is small enough, as you walk around $\partial \mathbf{R}(Q)$ in the direction of the described orientation with your left hand over $\mathbf{R}(Q)$, your head points roughly in the direction of $\mathbf{R}_u \times \mathbf{R}_v$.

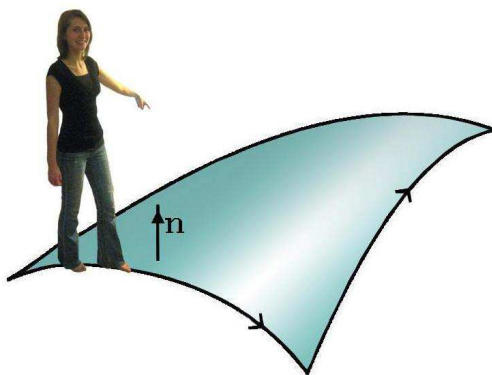


As explained above, this is true of the tangent parallelogram, and by continuity of $\mathbf{R}_v, \mathbf{R}_u$, the normals to the surface $\mathbf{R}(Q)$ $\mathbf{R}_u \times \mathbf{R}_v(\mathbf{u})$ for $\mathbf{u} \in Q$ will still point roughly in the same direction as your head if you walk in the indicated direction over $\partial \mathbf{R}(Q)$, meaning the angle between the vector from your feet to your head and the vector $\mathbf{R}_u \times \mathbf{R}_v(\mathbf{u})$ is less than $\pi/2$.

You can imagine filling U with such non-overlapping regions Q_i . Then orienting $\partial \mathbf{R}(Q_i)$ consistent with the counter clockwise orientation on Q_i , and adding the resulting line integrals, the line integrals over the common sides cancel as indicated in the following picture and the result is the line integral over ∂S .



Thus there is a simple relation between the field of normal vectors on S and the orientation of ∂S . It is simply this. If you walk along ∂S in the direction mandated by the orientation, with your left hand over the surface, the nearby normal vectors in Stoke's theorem will point roughly in the direction of your head.

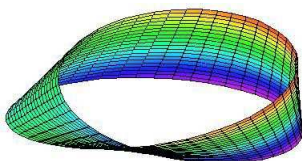


This also illustrates that you can **define** an orientation for ∂S by specifying a field of unit normal vectors for the surface, which varies continuously over the surface, and require that the motion over the boundary of the surface is such that your head points roughly in the direction of nearby normal vectors as you walk along the boundary with your left hand over S . The existence of such a continuous field of normal vectors is what constitutes an **orientable** surface.

30.3.2 The Mobeus Band

It turns out there are more general formulations of Stoke's theorem than what is presented above. However, it is always necessary for the surface S to be **orientable**. This means it is possible to obtain a vector field of unit normals to the surface which is a continuous function of position on S .

An example of a surface which is not orientable is the famous Mobeus band, obtained by taking a long rectangular piece of paper and gluing the ends together after putting a twist in it. Here is a picture of one.



There is something quite interesting about this Mobeus band and this is that it can be written parametrically with a simple parameter domain. The picture above is a maple graph of the parametrically defined surface

$$\mathbf{R}(\theta, v) \equiv \begin{cases} x = 4 \cos \theta + v \cos \frac{\theta}{2} \\ y = 4 \sin \theta + v \sin \frac{\theta}{2} \\ z = v \sin \frac{\theta}{2} \end{cases}, \quad \theta \in [0, 2\pi], v \in [-1, 1].$$

An obvious question is why the normal vector $\mathbf{R}_\theta \times \mathbf{R}_v / |\mathbf{R}_\theta \times \mathbf{R}_v|$ is not a continuous function of position on S . You can see easily that it is a continuous function of both θ and v . However, the map, \mathbf{R} is not one to one. In fact, $\mathbf{R}(0, 0) = \mathbf{R}(2\pi, 0)$. Therefore, near this point on S , there are two different values for the above normal vector. In fact, a tedious computation will show that this normal vector is

$$\frac{(4 \sin \frac{1}{2} \theta \cos \theta - \frac{1}{2} v, 4 \sin \frac{1}{2} \theta \sin \theta + \frac{1}{2} v, -8 \cos^2 \frac{1}{2} \theta \sin \frac{1}{2} \theta - 8 \cos^3 \frac{1}{2} \theta + 4 \cos \frac{1}{2} \theta)}{D}$$

where

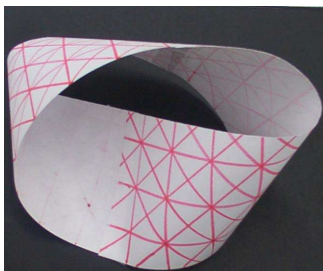
$$\begin{aligned} D = & 16 \sin^2 \left(\frac{\theta}{2} \right) + \frac{v^2}{2} + 4 \sin \left(\frac{\theta}{2} \right) v (\sin \theta - \cos \theta) \\ & + 4^3 \cos^2 \left(\frac{\theta}{2} \right) \left(\cos \left(\frac{1}{2} \theta \right) \sin \left(\frac{1}{2} \theta \right) + \cos^2 \left(\frac{1}{2} \theta \right) - \frac{1}{2} \right)^2 \end{aligned}$$

and you can verify that the denominator will not vanish. Letting $v = 0$ and $\theta = 0$ and 2π yields the two vectors $(0, 0, -1), (0, 0, 1)$ so there is a discontinuity. This is why I was careful to say in the statement of Stoke's theorem given above that \mathbf{R} is one to one.

The Mobeus band has some usefulness. In old machine shops the equipment was run by a belt which was given a twist to spread the surface wear on the belt over twice the area.

The above explanation shows that $\mathbf{R}_\theta \times \mathbf{R}_v / |\mathbf{R}_\theta \times \mathbf{R}_v|$ fails to deliver an orientation for the Mobeus band. However, this does not answer the question whether there is some orientation for it other than this one. In fact there is none. You can see this by looking at the first of the two pictures below or by making one and tracing it with a pencil. There is only one side to the Mobeus band. An oriented surface must have two sides, one side identified

by the given unit normal which varies continuously over the surface and the other side identified by the negative of this normal. The second picture below was taken by Ouyang when he was at meetings in Paris and saw it at a museum.

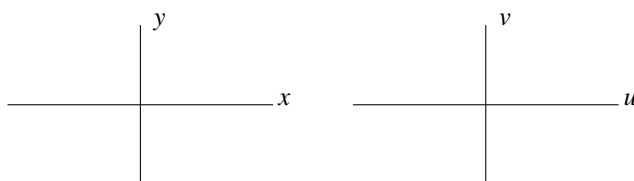


30.4 A General Green's Theorem

Now suppose U is a region in the uv plane for which Green's theorem holds and that

$$V \equiv \mathbf{R}(U)$$

where \mathbf{R} is $C^2(\overline{U}, \mathbb{R}^2)$ and is one to one, $\mathbf{R}_u \times \mathbf{R}_v \neq \mathbf{0}$. Here, to be specific, the u, v axes are oriented as the x, y axes respectively.



Also let $\mathbf{F}(x, y, z) = (P(x, y), Q(x, y), 0)$ be a C^1 vector field defined near V . Note that \mathbf{F} does not depend on z . Therefore,

$$\nabla \times \mathbf{F}(x, y) = (Q_x(x, y) - P_y(x, y)) \mathbf{k}.$$

You can check this from the definition. Also

$$\mathbf{R}(u, v) = \begin{pmatrix} x(u, v) \\ y(u, v) \end{pmatrix}$$

and so, from the definition of $\mathbf{R}_u \times \mathbf{R}_v$, the desired unit normal vector to V is

$$\frac{x_u y_v - x_v y_u}{|x_u y_v - x_v y_u|} \mathbf{k}$$

Suppose $x_u y_v - x_v y_u > 0$. Then the unit normal is \mathbf{k} . Then Stoke's theorem applied to this special case yields

$$\int_{\partial V} \mathbf{F} \cdot d\mathbf{R} = \int_U (Q_x(x(u, v), y(u, v)) - P_y(x(u, v), y(u, v))) \mathbf{k} \cdot \mathbf{k} \begin{vmatrix} x_u & x_v \\ y_u & y_v \end{vmatrix} dA$$

Now by the change of variables formula, this equals $\int_V (Q_x(x, y) - P_y(x, y)) dA$. This is just Green's theorem for V . Thus if U is a region for which Green's theorem holds and if V is

another region, $V = \mathbf{R}(U)$, where $|\mathbf{R}_u \times \mathbf{R}_v| \neq 0$, \mathbf{R} is one to one, and twice continuously differentiable with $\mathbf{R}_u \times \mathbf{R}_v$ in the direction of \mathbf{k} , then Green's theorem holds for V also.

This verifies the following theorem.

Theorem 30.4.1 (Green's Theorem) *Let V be an open set in the plane for which the divergence theorem holds and let ∂V be piecewise smooth and $\mathbf{F}(x, y) = (P(x, y), Q(x, y))$ be a C^1 vector field defined near V . Then if V is oriented counter clockwise,*

$$\int_{\partial V} \mathbf{F} \cdot d\mathbf{R} = \int_V \left(\frac{\partial Q}{\partial x}(x, y) - \frac{\partial P}{\partial y}(x, y) \right) dA. \quad (30.2)$$

In particular, if there exists U for which the divergence theorem holds and $V = \mathbf{R}(U)$ where $\mathbf{R}: U \rightarrow V$ is $C^2(\bar{U}, \mathbb{R}^2)$ such that $|\mathbf{R}_x \times \mathbf{R}_y| \neq 0$ and $\mathbf{R}_x \times \mathbf{R}_y$ is in the direction of \mathbf{k} , then 30.2 is valid where the orientation around ∂V is consistent with the orientation around U .

This is a very general version of Green's theorem which will include most if not all of what will be of interest. However, there are more general versions of this important theorem.¹

30.5 Conservative Vector Fields

Definition 30.5.1 *A vector field \mathbf{F} defined in a three dimensional region is said to be conservative² if for every piecewise smooth closed curve C , it follows $\int_C \mathbf{F} \cdot d\mathbf{R} = 0$.*

This looks a little different than the earlier definition. However, the main result in this section is an assertion that these are exactly the same.

Definition 30.5.2 *Let $(\mathbf{x}, \mathbf{p}_1, \dots, \mathbf{p}_n, \mathbf{y})$ be an ordered list of points in \mathbb{R}^p . Let $\mathbf{p}(\mathbf{x}, \mathbf{p}_1, \dots, \mathbf{p}_n, \mathbf{y})$ denote the piecewise smooth curve consisting of a straight line segment from \mathbf{x} to \mathbf{p}_1 and then the straight line segment from \mathbf{p}_1 to $\mathbf{p}_2 \dots$ and finally the straight line segment from \mathbf{p}_n to \mathbf{y} . This is called a **polygonal curve**. An open set in \mathbb{R}^p , U , is said to be a **region** if it has the property that for any two points $\mathbf{x}, \mathbf{y} \in U$, there exists a polygonal curve joining the two points.*

Conservative vector fields are important because of the following theorem, sometimes called the fundamental theorem for line integrals.

Theorem 30.5.3 *Let U be a region in \mathbb{R}^p and let $\mathbf{F}: U \rightarrow \mathbb{R}^p$ be a continuous vector field. Then \mathbf{F} is conservative if and only if there exists a scalar valued function of p variables ϕ such that $\mathbf{F} = \nabla \phi$. Furthermore, if C is an oriented curve which goes from \mathbf{x} to \mathbf{y} in U , then*

$$\int_C \mathbf{F} \cdot d\mathbf{R} = \phi(\mathbf{y}) - \phi(\mathbf{x}). \quad (30.3)$$

*Thus the line integral is path independent in this case. This function ϕ is called a **scalar potential** for \mathbf{F} .*

¹For a general version see the advanced calculus book by Apostol. Also see my book on calculus of real and complex variables. The general versions involve the concept of a rectifiable Jordan curve. You need to be able to take the area integral and to take the line integral around the boundary. This general version of this theorem appeared in 1951. Green lived in the early 1800's.

²There is no such thing as a liberal vector field.

Proof: To save space and fussing over things which are unimportant, denote by $p(x_0, x)$ a polygonal curve from x_0 to x . Thus the orientation is such that it goes from x_0 to x . The curve $p(x, x_0)$ denotes the same set of points but in the opposite order. Suppose first F is conservative. Fix $x_0 \in U$ and let

$$\phi(x) \equiv \int_{p(x_0, x)} F \cdot dR.$$

This is well defined because if $q(x_0, x)$ is another polygonal curve joining x_0 to x , Then the curve obtained by following $p(x_0, x)$ from x_0 to x and then from x to x_0 along $q(x, x_0)$ is a closed piecewise smooth curve and so by assumption, the line integral along this closed curve equals 0. However, this integral is just

$$\int_{p(x_0, x)} F \cdot dR + \int_{q(x, x_0)} F \cdot dR = \int_{p(x_0, x)} F \cdot dR - \int_{q(x_0, x)} F \cdot dR$$

which shows

$$\int_{p(x_0, x)} F \cdot dR = \int_{q(x_0, x)} F \cdot dR$$

and that ϕ is well defined. For small t ,

$$\begin{aligned} \frac{\phi(x + te_i) - \phi(x)}{t} &= \frac{\int_{p(x_0, x+te_i)} F \cdot dR - \int_{p(x_0, x)} F \cdot dR}{t} \\ &= \frac{\int_{p(x_0, x)} F \cdot dR + \int_{p(x, x+te_i)} F \cdot dR - \int_{p(x_0, x)} F \cdot dR}{t}. \end{aligned}$$

Since U is open, for small t , the ball of radius $|t|$ centered at x is contained in U . Therefore, the line segment from x to $x + te_i$ is also contained in U and so one can take

$$p(x, x + te_i)(s) = x + s(te_i)$$

for $s \in [0, 1]$. Therefore, the above difference quotient reduces to

$$\frac{1}{t} \int_0^1 F(x + s(te_i)) \cdot te_i ds = \int_0^1 F_i(x + s(te_i)) ds = F_i(x + s_t(te_i))$$

by the mean value theorem for integrals. Here s_t is some number between 0 and 1. By continuity of F , this converges to $F_i(x)$ as $t \rightarrow 0$. Therefore, $\nabla\phi = F$ as claimed.

Conversely, if $\nabla\phi = F$, then if $R: [a, b] \rightarrow \mathbb{R}^p$ is any C^1 curve joining x to y ,

$$\begin{aligned} \int_a^b F(R(t)) \cdot R'(t) dt &= \int_a^b \nabla\phi(R(t)) \cdot R'(t) dt = \int_a^b \frac{d}{dt}(\phi(R(t))) dt \\ &= \phi(R(b)) - \phi(R(a)) = \phi(y) - \phi(x) \end{aligned}$$

and this verifies 30.3 in the case where the curve joining the two points is smooth. The general case follows immediately from this by using this result on each of the pieces of the piecewise smooth curve. For example if the curve goes from x to p and then from p to y , the above would imply the integral over the curve from x to p is $\phi(p) - \phi(x)$ while from p to y the integral would yield $\phi(y) - \phi(p)$. Adding these gives $\phi(y) - \phi(x)$. The formula 30.3 implies the line integral over any closed curve equals zero because the starting and ending points of such a curve are the same. ■

Example 30.5.4 Let

$$\mathbf{F}(x, y, z) = (\cos x - yz \sin(xz), \cos(xz), -yx \sin(xz)).$$

Let C be a piecewise smooth curve which goes from $(\pi, 1, 1)$ to $(\frac{\pi}{2}, 3, 2)$. Find $\int_C \mathbf{F} \cdot d\mathbf{R}$.

The specifics of the curve are not given so the problem is nonsense unless the vector field is conservative. Therefore, it is reasonable to look for the function ϕ satisfying $\nabla\phi = \mathbf{F}$. Such a function satisfies $\phi_x = \cos x - y(\sin xz)z$ and so, assuming ϕ exists, $\phi(x, y, z) = \sin x + y \cos(xz) + \psi(y, z)$. I have to add in the most general thing possible, $\psi(y, z)$ to ensure possible solutions are not being thrown out. It wouldn't be good at this point to only add in a constant since the answer could involve a function of either or both of the other variables. Now from what was just obtained, $\phi_y = \cos(xz) + \psi_y = \cos xz$ and so it is possible to take $\psi_y = 0$. Consequently ϕ , if it exists, is of the form $\phi(x, y, z) = \sin x + y \cos(xz) + \psi(z)$. Now differentiating this with respect to z gives $\phi_z = -yx \sin(xz) + \psi_z = -yx \sin(xz)$ and this shows ψ does not depend on z either. Therefore, it suffices to take $\psi = 0$ and $\phi(x, y, z) = \sin(x) + y \cos(xz)$. Therefore, the desired line integral equals

$$\sin\left(\frac{\pi}{2}\right) + 3 \cos(\pi) - (\sin(\pi) + \cos(\pi)) = -1.$$

The above process for finding ϕ will not lead you astray in the case where there does not exist a scalar potential. As an example, consider the following.

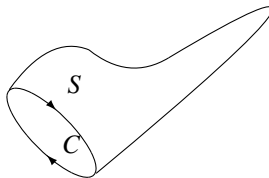
Example 30.5.5 Let $\mathbf{F}(x, y, z) = (x, y^2x, z)$. Find a scalar potential for \mathbf{F} if it exists.

If ϕ exists, then $\phi_x = x$ and so $\phi = \frac{x^2}{2} + \psi(y, z)$. Then $\phi_y = \psi_y(y, z) = xy^2$ but this is impossible because the left side depends only on y and z while the right side depends also on x . Therefore, this vector field is not conservative and there does not exist a scalar potential.

30.5.1 Some Terminology

If $\mathbf{F} = (P, Q, R)$ is a vector field. Then the statement that \mathbf{F} is conservative is the same as saying the differential form $Pdx + Qdy + Rdz$ is exact. Some people like to say things in terms of vector fields and some say it in terms of differential forms. In Example 30.5.8, the differential form $(4x^3 + 2(\cos(x^2 + z^2))x)dx + dy + (2(\cos(x^2 + z^2))z)dz$ is exact.

Definition 30.5.6 A set of points in three dimensional space V is simply connected if every piecewise smooth closed curve C is the edge of a surface S which is contained entirely within V in such a way that Stokes theorem holds for the surface S and its edge, C .



This is like a sock. The surface is the sock and the curve C goes around the opening of the sock.

As an application of Stoke's theorem, here is a useful theorem which gives a way to check whether a vector field is conservative.

Theorem 30.5.7 *For a three dimensional simply connected open set V and \mathbf{F} a C^1 vector field defined in V , \mathbf{F} is conservative if $\nabla \times \mathbf{F} = \mathbf{0}$ in V .*

Proof: If $\nabla \times \mathbf{F} = \mathbf{0}$ then taking an arbitrary closed curve C , and letting S be a surface bounded by C which is contained in V , Stoke's theorem implies

$$0 = \int_S \nabla \times \mathbf{F} \cdot \mathbf{n} dA = \int_C \mathbf{F} \cdot d\mathbf{R}.$$

Thus \mathbf{F} is conservative. ■

Example 30.5.8 *Determine whether the vector field*

$$(4x^3 + 2(\cos(x^2 + z^2))x, 1, 2(\cos(x^2 + z^2))z)$$

is conservative.

Since this vector field is defined on all of \mathbb{R}^3 , it only remains to take its curl and see if it is the zero vector.

$$\begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \partial_x & \partial_y & \partial_z \\ 4x^3 + 2(\cos(x^2 + z^2))x & 1 & 2(\cos(x^2 + z^2))z \end{vmatrix}.$$

This is obviously equal to zero. Therefore, the given vector field is conservative. Can you find a potential function for it? Let ϕ be the potential function. Then $\phi_z = 2(\cos(x^2 + z^2))z$ and so $\phi(x, y, z) = \sin(x^2 + z^2) + g(x, y)$. Now taking the derivative of ϕ with respect to y , you see $g_y = 1$ so $g(x, y) = y + h(x)$. Hence $\phi(x, y, z) = y + g(x) + \sin(x^2 + z^2)$. Taking the derivative with respect to x , you get $4x^3 + 2(\cos(x^2 + z^2))x = g'(x) + 2x\cos(x^2 + z^2)$ and so it suffices to take $g(x) = x^4$. Hence $\phi(x, y, z) = y + x^4 + \sin(x^2 + z^2)$.

30.6 Exercises

1. Determine whether the vector field

$$(2xy^3 \sin z^4, 3x^2y^2 \sin z^4 + 1, 4x^2y^3 (\cos z^4) z^3 + 1)$$

is conservative. If it is conservative, find a potential function.

2. Determine whether the vector field

$$(2xy^3 \sin z + y^2 + z, 3x^2y^2 \sin z + 2xy, x^2y^3 \cos z + x)$$

is conservative. If it is conservative, find a potential function.

3. Determine whether the vector field

$$(2xy^3 \sin z + z, 3x^2y^2 \sin z + 2xy, x^2y^3 \cos z + x)$$

is conservative. If it is conservative, find a potential function.

4. Find scalar potentials for the following vector fields if it is possible to do so. If it is not possible to do so, explain why.

(a) $(y^2, 2xy + \sin z, 2z + y \cos z)$

(b) $(2z(\cos(x^2 + y^2))x, 2z(\cos(x^2 + y^2))y, \sin(x^2 + y^2) + 2z)$

(c) $(f(x), g(y), h(z))$

(d) (xy, z^2, y^3)

(e) $\left(z + 2\frac{x}{x^2+y^2+1}, 2\frac{y}{x^2+y^2+1}, x + 3z^2\right)$

5. If a vector field is not conservative on the set U , is it possible the same vector field could be conservative on some subset of U ? Explain and give examples if it is possible. If it is not possible also explain why.

6. Prove that if a vector field \mathbf{F} has a scalar potential, then it has infinitely many scalar potentials.

7. Here is a vector field: $\mathbf{F} \equiv (2xy, x^2 - 5y^4, 3z^2)$. Find $\int_C \mathbf{F} \cdot d\mathbf{R}$ where C is a curve which goes from $(1, 2, 3)$ to $(4, -2, 1)$.

8. Here is a vector field: $\mathbf{F} \equiv (2xy, x^2 - 5y^4, 3(\cos z^3)z^2)$. Find $\int_C \mathbf{F} \cdot d\mathbf{R}$ where C is a curve which goes from $(1, 0, 1)$ to $(-4, -2, 1)$.

9. Find $\int_{\partial U} \mathbf{F} \cdot d\mathbf{R}$ where U is the set $\{(x, y) : 2 \leq x \leq 4, 0 \leq y \leq x\}$ and

$$\mathbf{F}(x, y) = (x \sin y, y \sin x)$$

10. Find $\int_{\partial U} \mathbf{F} \cdot d\mathbf{R}$ where U is $\{(x, y) : 2 \leq x \leq 3, 0 \leq y \leq x^2\}$ and

$$\mathbf{F}(x, y) = (x \cos y, y + x)$$

11. Find $\int_{\partial U} \mathbf{F} \cdot d\mathbf{R}$ where U is the set $\{(x, y) : 1 \leq x \leq 2, x \leq y \leq 3\}$ and

$$\mathbf{F}(x, y) = (x \sin y, y \sin x)$$

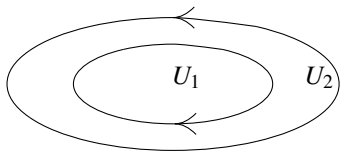
12. Find $\int_{\partial U} \mathbf{F} \cdot d\mathbf{R}$ where U is $\{(x, y) : x^2 + y^2 \leq 2\}$ and $\mathbf{F}(x, y) = (-y^3, x^3)$.

13. Show that for many open sets in \mathbb{R}^2 , Area of $U = \int_{\partial U} x dy$, and Area of $U = \int_{\partial U} -y dx$ and Area of $U = \frac{1}{2} \int_{\partial U} -y dx + x dy$. **Hint:** Use Green's theorem.

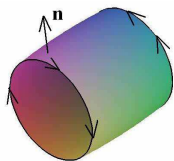
14. Two smooth oriented surfaces, S_1 and S_2 intersect in a smooth oriented closed curve C . Let \mathbf{F} be a C^1 vector field defined on \mathbb{R}^3 . Explain why $\int_{S_1} \text{curl}(\mathbf{F}) \cdot \mathbf{n} dS = \int_{S_2} \text{curl}(\mathbf{F}) \cdot \mathbf{n} dS$. Here \mathbf{n} is the normal to the surface which corresponds to the given orientation of the curve C .

15. Show that $\text{curl}(\psi \nabla \phi) = \nabla \psi \times \nabla \phi$ and explain why $\int_S \nabla \psi \times \nabla \phi \cdot \mathbf{n} dS = \int_{\partial S} (\psi \nabla \phi) \cdot d\mathbf{r}$.
16. Find a simple formula for $\text{div}(\nabla(u^\alpha))$ where $\alpha \in \mathbb{R}$.
17. Parametric equations for one arch of a cycloid are given by $x = a(t - \sin t)$ and $y = a(1 - \cos t)$ where here $t \in [0, 2\pi]$. Sketch a rough graph of this arch of a cycloid and then find the area between this arch and the x axis. **Hint:** This is very easy using Green's theorem and the vector field $\mathbf{F} = (-y, x)$.
18. Let $\mathbf{r}(t) = (\cos^3(t), \sin^3(t))$ where $t \in [0, 2\pi]$. Sketch this curve and find the area enclosed by it using Green's theorem.
19. Verify that Green's theorem can be considered a special case of Stoke's theorem.
20. Consider the vector field $\left(\frac{-y}{(x^2+y^2)}, \frac{x}{(x^2+y^2)}, 0 \right) = \mathbf{F}$. Show that $\nabla \times \mathbf{F} = \mathbf{0}$ but that for the closed curve, whose parametrization is $\mathbf{R}(t) = (\cos t, \sin t, 0)$ for $t \in [0, 2\pi]$, $\int_C \mathbf{F} \cdot d\mathbf{R} \neq 0$. Therefore, the vector field is not conservative. Does this contradict Theorem 30.5.7? Explain.
21. Let \mathbf{x} be a point of \mathbb{R}^3 and let \mathbf{n} be a unit vector. Let D_r be the circular disk of radius r containing \mathbf{x} which is perpendicular to \mathbf{n} . Placing the tail of \mathbf{n} at \mathbf{x} and viewing D_r from the point of \mathbf{n} , orient ∂D_r in the counter clockwise direction. Now suppose \mathbf{F} is a vector field defined near \mathbf{x} . Show that $\text{curl}(\mathbf{F}) \cdot \mathbf{n} = \lim_{r \rightarrow 0} \frac{1}{\pi r^2} \int_{\partial D_r} \mathbf{F} \cdot d\mathbf{R}$. This last integral is sometimes called the circulation density of \mathbf{F} . Explain how this shows that $\text{curl}(\mathbf{F}) \cdot \mathbf{n}$ measures the tendency for the vector field to "curl" around the point, the vector \mathbf{n} at the point \mathbf{x} .
22. The cylinder $x^2 + y^2 = 4$ is intersected with the plane $x + y + z = 2$. This yields a closed curve C . Orient this curve in the counter clockwise direction when viewed from a point high on the z axis. Let $\mathbf{F} = (x^2y, z + y, x^2)$. Find $\int_C \mathbf{F} \cdot d\mathbf{R}$.
23. The cylinder $x^2 + 4y^2 = 4$ is intersected with the plane $x + 3y + 2z = 1$. This yields a closed curve C . Orient this curve in the counter clockwise direction when viewed from a point high on the z axis. Let $\mathbf{F} = (y, z + y, x^2)$. Find $\int_C \mathbf{F} \cdot d\mathbf{R}$.
24. The cylinder $x^2 + y^2 = 4$ is intersected with the plane $x + 3y + 2z = 1$. This yields a closed curve C . Orient this curve in the clockwise direction when viewed from a point high on the z axis. Let $\mathbf{F} = (y, z + y, x)$. Find $\int_C \mathbf{F} \cdot d\mathbf{R}$.
25. Let $\mathbf{F} = (xz, z^2(y + \sin x), z^3y)$. Find the surface integral $\int_S \text{curl}(\mathbf{F}) \cdot \mathbf{n} dA$ where S is the surface $z = 4 - (x^2 + y^2)$, $z \geq 0$.
26. Let $\mathbf{F} = (xz, (y^3 + x), z^3y)$. Find the surface integral $\int_S \text{curl}(\mathbf{F}) \cdot \mathbf{n} dA$ where S is the surface $z = 16 - (x^2 + y^2)$, $z \geq 0$.
27. The cylinder $z = y^2$ intersects the surface $z = 8 - x^2 - 4y^2$ in a curve C which is oriented in the counter clockwise direction when viewed high on the z axis. Find $\int_C \mathbf{F} \cdot d\mathbf{R}$ if $\mathbf{F} = \left(\frac{z^2}{2}, xy, xz \right)$.

28. Tell which open sets are simply connected. The inside of a car radiator, A donut., The solid part of a cannon ball which contains a void on the interior. The inside of a donut which has had a **large** bite taken out of it, All of \mathbb{R}^3 except the z axis, All of \mathbb{R}^3 except the xy plane.
29. Let P be a polygon with vertices $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n), (x_1, y_1)$ encountered as you move over the boundary of the polygon which is assumed a simple closed curve in the counter clockwise direction. Using Problem 13, find a nice formula for the area of the polygon in terms of the vertices.
30. Here is a picture of two regions in the plane, U_1 and U_2 . Suppose Green's theorem holds for each of these regions. Explain why Green's theorem must also hold for the region which lies between them if the boundary is oriented as shown in the picture.



31. Here is a picture of a surface which has two bounding curves oriented as shown. Explain why Stoke's theorem will hold for such a surface and sketch a region in the plane which could serve as a parameter domain for this surface.



Theory of Linear Ordinary Differential Equations

32. The following is a short list of Laplace transforms. $f(t)$ denotes the function and $F(s)$ the Laplace transform. $f * g(t)$, the convolution, is given by

$$f * g(t) = \int_0^t f(t-u)g(u)du$$

Verify each of these formulas.

$f(t)$	$F(s)$	$f(t)$	$F(s)$	$f(t)$	$F(s)$
$t^n e^{at}$	$\frac{n!}{(s-a)^{n+1}}$	$t^n, n \in \mathbb{N}$	$\frac{n!}{s^{n+1}}$	$e^{at} \sin bt$	$\frac{b}{(s-a)^2 + b^2}$
$e^{at} \cos bt$	$\frac{s-a}{(s-a)^2 + b^2}$	$f * g(t)$	$F(s)G(s)$		

Using the table, explain why, for A an $n \times n$ matrix, there exists an $n \times n$ matrix $\Phi(t)$ satisfying

$$\mathcal{L}(\Phi)(s) = (sI - A)^{-1}$$

which has the property that all entries of the k^{th} derivative of $\Phi^{(k)}(t)$ have exponential growth. **Hint:** You should use the formula for the inverse of a matrix in terms of co-factors. The entries of $(sI - A)^{-1}$ will all be rational functions whose denominators can theoretically be factored into products of linear and irreducible quadratics. Thus each will be the Laplace transform of such a function just described. For the needed

theory of partial fractions, see Problem 40 on Page 48 and the following problem on that page.

33. \uparrow Letting $\Phi(t)$ be as in the above problem, explain why

$$\left(I - \frac{1}{s}A\right)^{-1} = \Phi(0) + \int_0^\infty e^{-ts}\Phi'(t) dt$$

Letting $s \rightarrow \infty$, explain why $\Phi(0) = I$. Next explain why $\Phi'(t) = A\Phi(t)$. For this last part, you might show that $A\Phi(t)$ and $\Phi'(t)$ have the same Laplace transform. Thus, they will be the same by Theorem 10.3.3. Thus

$$\Phi'(t) = A\Phi(t), \Phi(0) = I$$

This is called a fundamental matrix. Show there is at most one solution to the above initial value problem so it is **THE** fundamental matrix.

34. \uparrow Show the group property of this fundamental solution, that $\Phi(t+s) = \Phi(t)\Phi(s)$ for any $s, t \in \mathbb{R}$. Explain why $\Phi(-t) = \Phi(t)^{-1}$. **Hint:** Use Laplace transforms to show that if $\Psi'(t) = A\Psi(t)$, $\Psi(0) = 0$, then $\Psi(t) = 0$. Then consider for $s \in \mathbb{R}$, $\Psi(t) = \Phi(t+s) - \Phi(t)\Phi(s)$.

35. \uparrow Now show that there is exactly one solution to the initial value problem

$$\mathbf{x}'(t) = A\mathbf{x}(t) + \mathbf{f}(t), \mathbf{x}(0) = \mathbf{x}_0$$

and it is given by

$$\mathbf{x}(t) = \Phi(t)\mathbf{x}_0 + \Phi(t) \int_0^t \Phi(-s)\mathbf{f}(s) ds$$

You just did more than the entire mathematical substance of a typical course in undergraduate differential equations other than a few recipes for nonlinear equations. The above formula is called the variation of constants formula or Green's formula.

Chapter 31

Curvilinear Coordinates

31.1 Basis Vectors

In this chapter, I will use the repeated index summation convention unless stated otherwise. Thus, **a repeated index indicates a sum**. Also, it is helpful in order to keep things straight to always have the two repeated indices be on different levels. That is, I will write $a_i^j b_j$ and not $a_{ij} b_j$. The reason for this will become clear as the exposition proceeds.

The usual basis vectors are denoted by i, j, k and are as the following picture describes.

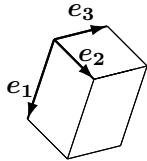


The vectors, i, j, k , are fixed. If v is a vector, there are unique scalars called components such that $v = v^1 i + v^2 j + v^3 k$. This is what it means that i, j, k is a basis. Review Section 18.4 at this time to see how this geometric notion relates to the general concept of a basis in a vector space.

Now suppose e_1, e_2, e_3 are three vectors which satisfy

$$e_1 \times e_2 \cdot e_3 \neq 0.$$

Recall this means the volume of the box spanned by the three vectors is not zero.



Suppose e_1, e_2, e_3 are as just described. Does it follow that they form a basis? In other words, for any vector v , there are unique scalars v^i such that $v = v^i e_i$. Of course this is the case because the box product is really the determinant of the matrix which has e_i as the i^{th} row (column). This is the content of the following theorem.

Theorem 31.1.1 *If e_1, e_2, e_3 are three vectors, then they form a basis if and only if*

$$e_1 \times e_2 \cdot e_3 \neq 0.$$

This gives a simple geometric condition which determines whether a list of three vectors forms a basis in \mathbb{R}^3 . One simply takes the box product. If the box product is not equal to zero, then the vectors form a basis. If not, the list of three vectors does not form a basis. This condition generalizes to \mathbb{R}^p as follows. If $e_i = a_i^j i_j$, then $\{e_i\}_{i=1}^p$ forms a basis if and only if $\det(a_i^j) \neq 0$.

These vectors may or may not be orthonormal. In any case, it is convenient to define something called the dual basis.

Definition 31.1.2 Let $\{e_i\}_{i=1}^p$ form a basis for \mathbb{R}^p . Then $\{e^i\}_{i=1}^p$ is called the dual basis if

$$e^i \cdot e_j = \delta_j^i \equiv \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}. \quad (31.1)$$

Theorem 31.1.3 If $\{e_i\}_{i=1}^p$ is a basis then $\{e^i\}_{i=1}^p$ is also a basis provided 31.1 holds.

Proof: Suppose

$$v = v_i e^i. \quad (31.2)$$

Then taking the dot product of both sides of 31.2 with e_j , yields

$$v_j = v \cdot e_j. \quad (31.3)$$

Thus there is at most one choice of scalars v_j such that $v = v_j e^j$ and it is given by 31.3.

$$(v - v \cdot e_j e^j) \cdot e_k = 0$$

and so, since $\{e_i\}_{i=1}^p$ is a basis,

$$(v - v \cdot e_j e^j) \cdot w = 0$$

for all vectors w . It follows $v - v \cdot e_j e^j = \mathbf{0}$ and this shows $\{e^i\}_{i=1}^p$ is a basis. ■

In the above argument are obtained formulas for the components of a vector v , v_i , with respect to the dual basis, found to be $v_j = v \cdot e_j$. In the same way, one can find the components of a vector with respect to the basis $\{e_i\}_{i=1}^p$. Let v be any vector and let

$$v = v^j e_j. \quad (31.4)$$

Then taking the dot product of both sides of 31.4 with e^i we see $v^i = e^i \cdot v$.

Does there exist a dual basis and is it uniquely determined?

Theorem 31.1.4 If $\{e_i\}_{i=1}^p$ is a basis for \mathbb{R}^p , then there exists a unique dual basis, $\{e^j\}_{j=1}^p$ satisfying

$$e^j \cdot e_i = \delta_i^j.$$

Proof: First I show the dual basis is unique. Suppose $\{f^j\}_{j=1}^p$ is another set of vectors which satisfies $f^j \cdot e_i = \delta_i^j$. Then

$$f^j = f^j \cdot e_i e^i = \delta_i^j e^i = e^j.$$

Note that from the definition, the dual basis to $\{e_j\}_{j=1}^p$ is just $e^j = e_j$. It remains to verify the existence of the dual basis. Consider the matrix $g_{ij} \equiv e_i \cdot e_j$. This is called the **metric tensor**. If the resulting matrix is denoted as G , does it follow that G^{-1} exists? Suppose you have $e_i \cdot e_j x^j = 0$. Then, since i is arbitrary, this implies $e_j x^j = \mathbf{0}$ and since $\{e_j\}$ is a basis, this requires each x^j to be zero. Thus G is invertible. Denote by g^{ij} the ij^{th} entry of this inverse matrix. Consider $e^j \equiv g^{jk} e_k$. Is this the dual basis as the notation implies?

$$e^j \cdot e_i = g^{jk} e_k \cdot e_i = g^{jk} g_{ki} = \delta_i^j$$

so yes, it is indeed the dual basis. This has shown both existence and uniqueness of the dual basis. ■

From this is a useful observation.

Proposition 31.1.5 $\{e_i\}_{i=1}^p$ is a basis for \mathbb{R}^p if and only if when $e_i = a_i^j \mathbf{i}_j$, $\det(a_i^j) \neq 0$.

Proof: First suppose $\{e_i\}_{i=1}^p$ is a basis for \mathbb{R}^p . Letting $A_{ij} \equiv a_i^j$, we need to show that $\det(A) \neq 0$. This is equivalent to showing that A or A^T is one to one. But

$$a_i^j x^j = 0 \Rightarrow a_i^j x^j \mathbf{i}_j = 0 \Rightarrow e_i x^i = 0 \Rightarrow x^i = 0$$

so A^T is one to one if and only if $\det(A) = \det(A^T) \neq 0$.

Conversely, suppose A has nonzero determinant. Why are the e_k a basis? Suppose $x^k e_k = \mathbf{0}$. Is each $x^k = 0$? Then $x^k a_k^j \mathbf{i}_j = \mathbf{0}$ and so for each j , $a_k^j x^k = 0$ and since A has nonzero determinant, $x^k = 0$. ■

Summarizing what has been shown so far, we know that $\{e_i\}_{i=1}^p$ is a basis for \mathbb{R}^p if and only if when $e_i = a_i^j \mathbf{i}_j$,

$$\det(a_i^j) \neq 0. \quad (31.5)$$

If $\{e_i\}_{i=1}^p$ is a basis, then there exists a unique dual basis, $\{e^j\}_{j=1}^p$ satisfying

$$e^j \cdot e_i = \delta_i^j, \quad (31.6)$$

and that if \mathbf{v} is any vector,

$$\mathbf{v} = v_j e^j, \quad \mathbf{v} = v^j e_j. \quad (31.7)$$

The components of \mathbf{v} which have the index on the top are called the contravariant components of the vector while the components which have the index on the bottom are called the covariant components. In general $v_i \neq v^i$! We also have formulae for these components in terms of the dot product.

$$v_j = \mathbf{v} \cdot e_j, \quad v^j = \mathbf{v} \cdot e^j. \quad (31.8)$$

As indicated above, define $g_{ij} \equiv e_i \cdot e_j$ and $g^{ij} \equiv e^i \cdot e^j$. The next theorem describes the process of raising or lowering an index.

Theorem 31.1.6 *The following hold.*

$$g^{ij} e_j = e^i, \quad g_{ij} e^j = e_i, \quad (31.9)$$

$$g^{ij} v_j = v^i, \quad g_{ij} v^j = v_i, \quad (31.10)$$

$$g^{ij} g_{jk} = \delta_k^i, \quad (31.11)$$

$$\det(g_{ij}) > 0, \quad \det(g^{ij}) > 0. \quad (31.12)$$

Proof: First,

$$e^i = e^i \cdot e^j e_j = g^{ij} e_j$$

by 31.7 and 31.8. Similarly, by 31.7 and 31.8,

$$e_i = e_i \cdot e^j e_j = g_{ij} e^j.$$

This verifies 31.9. To verify 31.10,

$$v^i = e^i \cdot \mathbf{v} = g^{ij} e_j \cdot \mathbf{v} = g^{ij} v_j.$$

The proof of the remaining formula in 31.10 is similar.

To verify 31.11,

$$g^{ij}g_{jk} = \mathbf{e}^i \cdot \mathbf{e}^j \mathbf{e}_j \cdot \mathbf{e}_k = ((\mathbf{e}^i \cdot \mathbf{e}^j) \mathbf{e}_j) \cdot \mathbf{e}_k = \mathbf{e}^i \cdot \mathbf{e}_k = \delta_k^i.$$

This shows the two determinants in 31.12 are non zero because the two matrices are inverses of each other. It only remains to verify that one of these is greater than zero. Letting $\mathbf{e}_i = a_i^j \mathbf{i}_j = b_j^i \mathbf{i}^j$, we see that since $\mathbf{i}_j = \mathbf{i}^j$, $a_i^j = b_j^i$. Therefore,

$$\mathbf{e}_i \cdot \mathbf{e}_j = a_i^r \mathbf{i}_r \cdot b_k^j \mathbf{i}^k = a_i^r b_k^j \delta_r^k = a_i^k b_k^j = a_i^k a_j^k.$$

It follows that for G the matrix whose ij^{th} entry is $\mathbf{e}_i \cdot \mathbf{e}_j$, $G = AA^T$ where the ik^{th} entry of A is a_i^k . Therefore, $\det(G) = \det(A) \det(A^T) = \det(A)^2 > 0$. It follows from 31.11 that if H is the matrix whose ij^{th} entry is g^{ij} , then $GH = I$ and so $H = G^{-1}$ and

$$\det(G) \det(G^{-1}) = \det(g^{ij}) \det(G) = 1.$$

Therefore, $\det(G^{-1}) > 0$ also. ■

Note that $\det(AA^T) \geq 0$ always, because the eigenvalues are nonnegative.

As noted above, we have the following definition.

Definition 31.1.7 The matrix $(g_{ij}) = G$ is called the metric tensor.

31.2 Exercises

1. Let $\mathbf{e}_1 = \mathbf{i} + \mathbf{j}$, $\mathbf{e}_2 = \mathbf{i} - \mathbf{j}$, $\mathbf{e}_3 = \mathbf{j} + \mathbf{k}$. Find $\mathbf{e}^1, \mathbf{e}^2, \mathbf{e}^3$, (g_{ij}) , (g^{ij}) . If $\mathbf{v} = \mathbf{i} + 2\mathbf{j} + \mathbf{k}$, find v^i and v_j , the contravariant and covariant components of the vector.
2. Let $\mathbf{e}^1 = 2\mathbf{i} + \mathbf{j}$, $\mathbf{e}^2 = \mathbf{i} - 2\mathbf{j}$, $\mathbf{e}^3 = \mathbf{k}$. Find $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$, (g_{ij}) , (g^{ij}) . If $\mathbf{v} = 2\mathbf{i} - 2\mathbf{j} + \mathbf{k}$, find v^i and v_j , the contravariant and covariant components of the vector.
3. Suppose $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ have the property that $\mathbf{e}_i \cdot \mathbf{e}_j = 0$ whenever $i \neq j$. Show the same is true of the dual basis.
4. Let $\mathbf{e}_1, \dots, \mathbf{e}_3$ be a basis for \mathbb{R}^n and let $\mathbf{v} = v^i \mathbf{e}_i = v_i \mathbf{e}^i$, $\mathbf{w} = w^j \mathbf{e}_j = w_j \mathbf{e}^j$ be two vectors. Show

$$\mathbf{v} \cdot \mathbf{w} = g_{ij} v^i w^j = g^{ij} v_i w_j.$$

5. Show if $\{\mathbf{e}_i\}_{i=1}^3$ is a basis in \mathbb{R}^3

$$\mathbf{e}^1 = \frac{\mathbf{e}_2 \times \mathbf{e}_3}{\mathbf{e}_2 \times \mathbf{e}_3 \cdot \mathbf{e}_1}, \mathbf{e}^2 = \frac{\mathbf{e}_1 \times \mathbf{e}_3}{\mathbf{e}_1 \times \mathbf{e}_3 \cdot \mathbf{e}_2}, \mathbf{e}^3 = \frac{\mathbf{e}_1 \times \mathbf{e}_2}{\mathbf{e}_1 \times \mathbf{e}_2 \cdot \mathbf{e}_3}.$$

6. Let $\{\mathbf{e}_i\}_{i=1}^n$ be a basis and define

$$\mathbf{e}_i^* \equiv \frac{\mathbf{e}_i}{|\mathbf{e}_i|}, \mathbf{e}^{*i} \equiv \mathbf{e}^i |\mathbf{e}_i|.$$

Show $\mathbf{e}^{*i} \cdot \mathbf{e}_j^* = \delta_j^i$.

7. If \mathbf{v} is a vector, v_i^* and v^{*i} , are defined by

$$\mathbf{v} \equiv v_i^* \mathbf{e}^{*i} \equiv v^{*i} \mathbf{e}_i^*.$$

These are called the physical components of \mathbf{v} . Show

$$v_i^* = \frac{v_i}{|\mathbf{e}_i|}, \quad v^{*i} = v^i |\mathbf{e}_i| \quad (\text{No summation on } i).$$

31.3 Curvilinear Coordinates

There are many ways to identify a point in n dimensional space with an ordered list of real numbers. Some of these are spherical coordinates, cylindrical coordinates and rectangular coordinates and these particular examples are discussed earlier. I will denote by \mathbf{y} the rectangular coordinates of a point in n dimensional space which I will go on writing as \mathbb{R}^n . Thus $\mathbf{y} = (y^1 \cdots y^n)$. It follows there are equations which relate the rectangular coordinates to some other coordinates $(x^1 \cdots x^n)$. In spherical coordinates, these were ρ, ϕ, θ where the geometric meaning of these were described earlier. However, completely general systems are to be considered here, with certain stipulations. The idea is

$$y^k = y^k(x^1, \dots, x^n), \quad \mathbf{y} = \mathbf{y}(x^1, \dots, x^n)$$

Let $(x^1 \cdots x^n) \in D \subseteq \mathbb{R}^n$ be an open set and let $\mathbf{x} \rightarrow \mathbf{y}(x^1, \dots, x^n) \equiv \mathbf{M}(x^1, \dots, x^n)$ satisfy

$$\mathbf{M} \text{ is } C^2, \quad (31.13)$$

$$\mathbf{M} \text{ is one to one.} \quad (31.14)$$

Letting $\mathbf{x} \in D$, we can write

$$\mathbf{M}(\mathbf{x}) = M^k(\mathbf{x}) \mathbf{i}_k$$

where, as usual, \mathbf{i}_k are the standard basis vectors for \mathbb{R}^n , \mathbf{i}_k being the vector in \mathbb{R}^n which has a one in the k^{th} coordinate and a 0 in every other spot. Thus $y^k = M^k(\mathbf{x})$ where this y^k refers to the k^{th} rectangular coordinate of the point \mathbf{y} as just described.

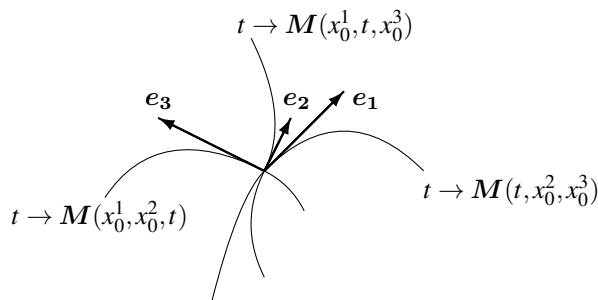
For a fixed $\mathbf{x} \in D$, we can consider the space curves,

$$t \rightarrow \mathbf{M}(\mathbf{x} + t\mathbf{i}_k) \equiv \mathbf{y}(\mathbf{x} + t\mathbf{i}_k)$$

for $t \in I$, some open interval containing 0. Then for the point \mathbf{x} , we let

$$\mathbf{e}_k \equiv \frac{\partial \mathbf{M}}{\partial x^k}(\mathbf{x}) \equiv \frac{d}{dt}(\mathbf{M}(\mathbf{x} + t\mathbf{i}_k))|_{t=0} \equiv \frac{\partial \mathbf{y}}{\partial x^k}(\mathbf{x})$$

Denote this vector as $\mathbf{e}_k(\mathbf{x})$ to emphasize its dependence on \mathbf{x} . The following picture illustrates the situation in \mathbb{R}^3 .



I want $\{e_k\}_{k=1}^n$ to be a basis. Thus, from Proposition 31.1.5,

$$\det\left(\frac{\partial M^i}{\partial x^k}\right) \equiv \det(D\mathbf{y}(x)) \equiv \det(D(M)(x)) \neq 0. \quad (31.15)$$

Let

$$y^i = M^i(x) \quad i = 1, \dots, n \quad (31.16)$$

so that the y^i are the usual rectangular coordinates with respect to the usual basis vectors $\{i_k\}_{k=1}^n$ of the point $\mathbf{y} = M(x)$. Letting $x \equiv (x^1, \dots, x^n)$, it follows from the inverse function theorem (See Chapter 24) that $M(D)$ is open, and that 31.15, 31.13, and 31.14 imply the equations 31.16 define each x^i as a C^2 function of $\mathbf{y} \equiv (y^1, \dots, y^n)$. Thus, abusing notation slightly, the equations 31.16 are equivalent to

$$x^i = x^i(y^1, \dots, y^n), \quad i = 1, \dots, n$$

where x^i is a C^2 function of the rectangular coordinates of a point \mathbf{y} . It follows from the material on the gradient described earlier,

$$\nabla x^k(\mathbf{y}) = \frac{\partial x^k(\mathbf{y})}{\partial y^j} i^j.$$

Then

$$\nabla x^k(\mathbf{y}) \cdot e_j = \frac{\partial x^k}{\partial y^s} i^s \cdot \frac{\partial y^r}{\partial x^j} i_r = \frac{\partial x^k}{\partial y^s} \frac{\partial y^s}{\partial x^j} = \delta_j^k$$

by the chain rule. Therefore, the dual basis is given by

$$e^k(x) = \nabla x^k(\mathbf{y}(x)). \quad (31.17)$$

Notice that it might be hard or even impossible to solve algebraically for x^i in terms of the y^j . Thus the straight forward approach to finding e^k by 31.17 might be impossible. Also, this approach leads to an expression in terms of the \mathbf{y} coordinates rather than the desired \mathbf{x} coordinates. Therefore, it is expedient to use another method to obtain these vectors in terms of \mathbf{x} . Indeed, this is the main idea in this chapter, doing everything in terms of \mathbf{x} rather than \mathbf{y} . The vectors, $e^k(x)$ may always be found by using formula 31.9 and the result is in terms of the curvilinear coordinates \mathbf{x} . Here is a familiar example.

Example 31.3.1 $D \equiv (0, \infty) \times (0, \pi) \times (0, 2\pi)$ and

$$\begin{pmatrix} y^1 \\ y^2 \\ y^3 \end{pmatrix} = \begin{pmatrix} x^1 \sin(x^2) \cos(x^3) \\ x^1 \sin(x^2) \sin(x^3) \\ x^1 \cos(x^2) \end{pmatrix}$$

(We usually write this as

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \rho \sin(\phi) \cos(\theta) \\ \rho \sin(\phi) \sin(\theta) \\ \rho \cos(\phi) \end{pmatrix}$$

where (ρ, ϕ, θ) are the spherical coordinates. We are calling them x^1, x^2 , and x^3 to preserve the notation just discussed.) Thus

$$\mathbf{e}_1(\mathbf{x}) = \sin(x^2) \cos(x^3) \mathbf{i}_1 + \sin(x^2) \sin(x^3) \mathbf{i}_2 + \cos(x^2) \mathbf{i}_3,$$

$$\mathbf{e}_2(\mathbf{x}) = x^1 \cos(x^2) \cos(x^3) \mathbf{i}_1$$

$$+ x^1 \cos(x^2) \sin(x^3) \mathbf{i}_2 - x^1 \sin(x^2) \mathbf{i}_3,$$

$$\mathbf{e}_3(\mathbf{x}) = -x^1 \sin(x^2) \sin(x^3) \mathbf{i}_1 + x^1 \sin(x^2) \cos(x^3) \mathbf{i}_2 + 0 \mathbf{i}_3.$$

It follows the metric tensor is

$$G = \begin{pmatrix} 1 & 0 & 0 \\ 0 & (x^1)^2 & 0 \\ 0 & 0 & (x^1)^2 \sin^2(x^2) \end{pmatrix} = (g_{ij}) = (\mathbf{e}_i \cdot \mathbf{e}_j). \quad (31.18)$$

Therefore, by Theorem 31.1.6

$$\begin{aligned} G^{-1} &= (g^{ij}) \\ &= (\mathbf{e}^i, \mathbf{e}^j) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & (x^1)^{-2} & 0 \\ 0 & 0 & (x^1)^{-2} \sin^{-2}(x^2) \end{pmatrix}. \end{aligned}$$

To obtain the dual basis, use Theorem 31.1.6 to write

$$\mathbf{e}^1(\mathbf{x}) = g^{1j} \mathbf{e}_j(\mathbf{x}) = \mathbf{e}_1(\mathbf{x})$$

$$\mathbf{e}^2(\mathbf{x}) = g^{2j} \mathbf{e}_j(\mathbf{x}) = (x^1)^{-2} \mathbf{e}_2(\mathbf{x})$$

$$\mathbf{e}^3(\mathbf{x}) = g^{3j} \mathbf{e}_j(\mathbf{x}) = (x^1)^{-2} \sin^{-2}(x^2) \mathbf{e}_3(\mathbf{x}).$$

Note that $\frac{\partial \mathbf{y}}{\partial y^k} \equiv \mathbf{e}_k(\mathbf{y}) = \mathbf{i}^k = \mathbf{i}_k$ where, as described, $(y^1 \cdots y^n)$ are the rectangular coordinates of the point in \mathbb{R}^n .

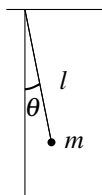
31.4 Exercises

1. Let

$$\begin{pmatrix} y^1 \\ y^2 \\ y^3 \end{pmatrix} = \begin{pmatrix} x^1 + 2x^2 \\ x^2 + x^3 \\ x^1 - 2x^2 \end{pmatrix}$$

where the y^i are the rectangular coordinates of the point. Find $e^i, e_i, i = 1, 2, 3$, and find $(g_{ij})(x)$ and $(g^{ij})(x)$.

2. Let $\mathbf{y} = \mathbf{y}(x, t)$ where t signifies time and $\mathbf{x} \in U \subseteq \mathbb{R}^m$ for U an open set, while $\mathbf{y} \in \mathbb{R}^n$ and suppose \mathbf{x} is a function of t . Physically, this corresponds to an object moving over a surface in \mathbb{R}^n which may be changing as a function of t . The point $\mathbf{y} = \mathbf{y}(x(t), t)$ is the point in \mathbb{R}^n corresponding to t . For example, consider the pendulum



in which $n = 2, l$ is fixed and $y^1 = l \sin \theta, y^2 = l - l \cos \theta$. Thus, in this simple example, $m = 1$. If l were changing in a known way with respect to t , then this would be of the form $\mathbf{y} = \mathbf{y}(x, t)$. In general, the kinetic energy is defined as

$$T \equiv \frac{1}{2} m \dot{\mathbf{y}} \cdot \dot{\mathbf{y}} \quad (*)$$

where the dot on the top signifies differentiation with respect to t . Show

$$\frac{\partial T}{\partial \dot{x}^k} = m \dot{\mathbf{y}} \cdot \frac{\partial \mathbf{y}}{\partial x^k}.$$

Hint: First show

$$\dot{\mathbf{y}} = \frac{\partial \mathbf{y}}{\partial x^j} \dot{x}^j + \frac{\partial \mathbf{y}}{\partial t} \quad (**)$$

and so

$$\frac{\partial \dot{\mathbf{y}}}{\partial \dot{x}^j} = \frac{\partial \mathbf{y}}{\partial x^j}.$$

3. \uparrow Show

$$\frac{d}{dt} \left(\frac{\partial T}{\partial \dot{x}^k} \right) = m \ddot{\mathbf{y}} \cdot \frac{\partial \mathbf{y}}{\partial x^k} + m \dot{\mathbf{y}} \cdot \frac{\partial^2 \mathbf{y}}{\partial x^k \partial x^r} \dot{x}^r + m \dot{\mathbf{y}} \cdot \frac{\partial^2 \mathbf{y}}{\partial t \partial x^k}.$$

4. \uparrow Show

$$\frac{\partial T}{\partial x^k} = m \dot{\mathbf{y}} \cdot \left(\frac{\partial^2 \mathbf{y}}{\partial x^r \partial x^k} \dot{x}^r + \frac{\partial^2 \mathbf{y}}{\partial t \partial x^k} \right).$$

Hint: Use $*$ and $**$.

5. ↑ Now show from Newton's second law (mass times acceleration equals force) that for \mathbf{F} the force,

$$\frac{d}{dt} \left(\frac{\partial T}{\partial \dot{x}^k} \right) - \frac{\partial T}{\partial x^k} = m \ddot{\mathbf{y}} \cdot \frac{\partial \mathbf{y}}{\partial x^k} = \mathbf{F} \cdot \frac{\partial \mathbf{y}}{\partial x^k}. \quad (***)$$

6. ↑ In the example of the simple pendulum above,

$$\mathbf{y} = \begin{pmatrix} l \sin \theta \\ l - l \cos \theta \end{pmatrix} = l \sin \theta \mathbf{i} + (l - l \cos \theta) \mathbf{j}.$$

Use *** to find a differential equation which describes the vibrations of the pendulum in terms of θ . First write the kinetic energy and then consider the force acting on the mass which is $-mg\mathbf{j}$.

7. Of course, the idea is to write equations of motion in terms of the variables x^k , instead of the rectangular variables y^k . Suppose $\mathbf{y} = \mathbf{y}(x)$ and x is a function of t . Letting G denote the metric tensor, show that the kinetic energy is of the form $\frac{1}{2} m \dot{x}^T G x$ where m is a point mass with m its mass.
8. The pendulum problem is fairly easy to do without the formalism developed. Now consider the case where $x = (\rho, \theta, \phi)$, spherical coordinates, and write differential equations for ρ , θ , and ϕ to describe the motion of an object in terms of these coordinates given a force, \mathbf{F} .
9. Suppose the pendulum is not assumed to vibrate in a plane. Let it be suspended at the origin and let ϕ be the angle between the negative z axis and the positive x axis while θ is the angle between the projection of the position vector onto the xy plane and the positive x axis in the usual way. Thus

$$x = \rho \sin \phi \cos \theta, y = \rho \sin \phi \sin \theta, z = -\rho \cos \phi$$

10. If there are many masses, $m_\alpha, \alpha = 1, \dots, R$, the kinetic energy is the sum of the kinetic energies of the individual masses. Thus,

$$T \equiv \frac{1}{2} \sum_{\alpha=1}^R m_\alpha |\dot{\mathbf{y}}_\alpha|^2.$$

Generalize the above problems to show that, assuming

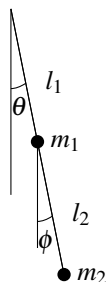
$$\mathbf{y}_\alpha = \mathbf{y}_\alpha(x, t),$$

$$\frac{d}{dt} \left(\frac{\partial T}{\partial \dot{x}^k} \right) - \frac{\partial T}{\partial x^k} = \sum_{\alpha=1}^R \mathbf{F}_\alpha \cdot \frac{\partial \mathbf{y}_\alpha}{\partial x^k}$$

where \mathbf{F}_α is the force acting on m_α .

11. Discuss the equivalence of these formulae with Newton's second law, force equals mass times acceleration. What is gained from the above so called Lagrangian formalism?

12. The double pendulum has two masses instead of only one.

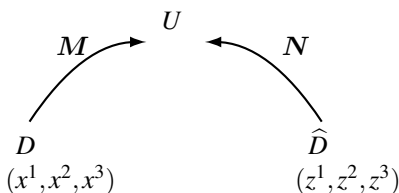


Write differential equations for θ and ϕ to describe the motion of the double pendulum.

31.5 Transformation of Coordinates.

How do we write $e^k(x)$ in terms of the vectors, $e^j(z)$ where z is some other type of curvilinear coordinates? This is next.

Consider the following picture in which U is an open set in \mathbb{R}^n , D and \hat{D} are open sets in \mathbb{R}^n , and M, N are C^2 mappings which are one to one from D and \hat{D} respectively. The only reason for this is to ensure that the mixed partial derivatives are equal. We will suppose that a point in U is identified by the curvilinear coordinates x in D and z in \hat{D} .



Thus $M(x) = N(z)$ and so $z = N^{-1}(M(x))$. The point in U will be denoted in rectangular coordinates as y and we have $y(x) = y(z)$. Now by the chain rule,

$$e_i(z) = \frac{\partial y}{\partial z^i} = \frac{\partial y}{\partial x^j} \frac{\partial x^j}{\partial z^i} = \frac{\partial x^j}{\partial z^i} e_j(x) \quad (31.19)$$

Define the covariant and contravariant coordinates for the various curvilinear coordinates in the obvious way. Thus,

$$v = v_i(x) e^i(x) = v^j(x) e_j(x) = v_j(z) e^j(z) = v^j(z) e_j(z).$$

Then the following theorem tells how to transform the vectors and coordinates.

Theorem 31.5.1 *The following transformation rules hold for pairs of curvilinear coordinates.*

$$v_i(z) = \frac{\partial x^j}{\partial z^i} v_j(x), \quad v^i(z) = \frac{\partial z^i}{\partial x^j} v^j(x), \quad (31.20)$$

$$e_i(z) = \frac{\partial x^j}{\partial z^i} e_j(x), \quad e^i(z) = \frac{\partial z^i}{\partial x^j} e^j(x), \quad (31.21)$$

$$g_{ij}(z) = \frac{\partial x^r}{\partial z^i} \frac{\partial x^s}{\partial z^j} g_{rs}(x), \quad g^{ij}(z) = \frac{\partial z^i}{\partial x^r} \frac{\partial z^j}{\partial x^s} g^{rs}(x). \quad (31.22)$$

Proof: We already have shown the first part of 31.21 in 31.19. Then, from 31.19,

$$\begin{aligned} e^i(z) &= e^i(z) \cdot e_j(x) e^j(x) = e^i(z) \cdot \frac{\partial z^k}{\partial x^j} e_k(z) e^j(x) \\ &= \delta_k^i \frac{\partial z^k}{\partial x^j} e^j(x) = \frac{\partial z^i}{\partial x^j} e^j(x) \end{aligned}$$

and this proves the second part of 31.21. Now to show 31.20,

$$v_i(z) = v \cdot e_i(z) = v \cdot \frac{\partial x^j}{\partial z^i} e_j(x) = \frac{\partial x^j}{\partial z^i} v \cdot e_j(x) = \frac{\partial x^j}{\partial z^i} v_j(x)$$

and

$$v^i(z) = v \cdot e^i(z) = v \cdot \frac{\partial z^i}{\partial x^j} e^j(x) = \frac{\partial z^i}{\partial x^j} v \cdot e^j(x) = \frac{\partial z^i}{\partial x^j} v^j(x).$$

To verify 31.22,

$$g_{ij}(z) = e_i(z) \cdot e_j(z) = e_r(x) \frac{\partial x^r}{\partial z^i} \cdot e_s(x) \frac{\partial x^s}{\partial z^j} = g_{rs}(x) \frac{\partial x^r}{\partial z^i} \frac{\partial x^s}{\partial z^j}. \blacksquare$$

31.6 Differentiation and Christoffel Symbols

Let $\mathbf{F} : U \rightarrow \mathbb{R}^n$ be differentiable. We call \mathbf{F} a vector field and it is used to model force, velocity, acceleration, or any other vector quantity which may change from point to point in U . Then $\frac{\partial \mathbf{F}(\mathbf{x})}{\partial x^j}$ is a vector and so there exist scalars, $F_{,j}^i(\mathbf{x})$ and $F_{i,j}(\mathbf{x})$ such that

$$\frac{\partial \mathbf{F}(\mathbf{x})}{\partial x^j} = F_{,j}^i(\mathbf{x}) e_i(\mathbf{x}), \quad \frac{\partial \mathbf{F}(\mathbf{x})}{\partial x^j} = F_{i,j}(\mathbf{x}) e^i(\mathbf{x}) \quad (31.23)$$

We will see how these scalars transform when the coordinates are changed.

Theorem 31.6.1 *If \mathbf{x} and \mathbf{z} are curvilinear coordinates,*

$$F_{,s}^r(\mathbf{x}) = F_{,j}^i(\mathbf{z}) \frac{\partial x^r}{\partial z^i} \frac{\partial z^j}{\partial x^s}, \quad F_{r,s}(\mathbf{x}) \frac{\partial x^r}{\partial z^i} \frac{\partial x^s}{\partial z^j} = F_{i,j}(\mathbf{z}). \quad (31.24)$$

Proof:

$$\begin{aligned} F_{,s}^r(\mathbf{x}) e_r(\mathbf{x}) &\equiv \frac{\partial \mathbf{F}(\mathbf{x})}{\partial x^s} = \frac{\partial \mathbf{F}(\mathbf{z})}{\partial z^j} \frac{\partial z^j}{\partial x^s} \equiv \\ &F_{,j}^i(\mathbf{z}) e_i(\mathbf{z}) \frac{\partial z^j}{\partial x^s} = F_{,j}^i(\mathbf{z}) \frac{\partial z^j}{\partial x^s} \frac{\partial x^r}{\partial z^i} e_r(\mathbf{x}) \end{aligned}$$

which shows the first formula of 31.23. To show the other formula,

$$\begin{aligned} F_{i,j}(\mathbf{z}) e^i(\mathbf{z}) &\equiv \frac{\partial \mathbf{F}(\mathbf{z})}{\partial z^j} = \frac{\partial \mathbf{F}(\mathbf{x})}{\partial x^s} \frac{\partial x^s}{\partial z^j} \equiv \\ &F_{r,s}(\mathbf{x}) e^r(\mathbf{x}) \frac{\partial x^s}{\partial z^j} = F_{r,s}(\mathbf{x}) \frac{\partial x^s}{\partial z^j} \frac{\partial x^r}{\partial z^i} e^i(\mathbf{z}), \end{aligned}$$

and this shows the second formula for transforming these scalars. ■

Now $\mathbf{F}(\mathbf{x}) = F^i(\mathbf{x}) \mathbf{e}_i(\mathbf{x})$ and so by the product rule,

$$\frac{\partial \mathbf{F}}{\partial x^j} = \frac{\partial F^i}{\partial x^j} \mathbf{e}_i(\mathbf{x}) + F^i(\mathbf{x}) \frac{\partial \mathbf{e}_i(\mathbf{x})}{\partial x^j}. \quad (31.25)$$

Now $\frac{\partial \mathbf{e}_i(\mathbf{x})}{\partial x^j}$ is a vector and so there exist scalars, $\left\{ \begin{smallmatrix} k \\ ij \end{smallmatrix} \right\}$ such that

$$\frac{\partial \mathbf{e}_i(\mathbf{x})}{\partial x^j} = \left\{ \begin{smallmatrix} k \\ ij \end{smallmatrix} \right\} \mathbf{e}_k(\mathbf{x}).$$

Thus

$$\left\{ \begin{smallmatrix} k \\ ij \end{smallmatrix} \right\} \mathbf{e}_k(\mathbf{x}) = \frac{\partial^2 \mathbf{y}}{\partial x^j \partial x^i}$$

and so

$$\left\{ \begin{smallmatrix} k \\ ij \end{smallmatrix} \right\} \mathbf{e}_k(\mathbf{x}) \cdot \mathbf{e}^r(\mathbf{x}) = \left\{ \begin{smallmatrix} k \\ ij \end{smallmatrix} \right\} \delta_k^r = \left\{ \begin{smallmatrix} r \\ ij \end{smallmatrix} \right\} = \frac{\partial^2 \mathbf{y}}{\partial x^j \partial x^i} \cdot \mathbf{e}^r(\mathbf{x}) \quad (31.26)$$

Therefore, from 31.25, $\frac{\partial \mathbf{F}}{\partial x^j} = \frac{\partial F^k}{\partial x^j} \mathbf{e}_k(\mathbf{x}) + F^i(\mathbf{x}) \left\{ \begin{smallmatrix} r \\ ij \end{smallmatrix} \right\} \mathbf{e}_k(\mathbf{x})$ which shows

$$F_{,j}^k(\mathbf{x}) = \frac{\partial F^k}{\partial x^j} + F^i(\mathbf{x}) \left\{ \begin{smallmatrix} k \\ ij \end{smallmatrix} \right\}. \quad (31.27)$$

This is sometimes called the covariant derivative.

Theorem 31.6.2 *The Christoffel symbols of the second kind satisfy the following*

$$\frac{\partial \mathbf{e}_i(\mathbf{x})}{\partial x^j} = \left\{ \begin{smallmatrix} k \\ ij \end{smallmatrix} \right\} \mathbf{e}_k(\mathbf{x}), \quad (31.28)$$

$$\frac{\partial \mathbf{e}^i(\mathbf{x})}{\partial x^j} = - \left\{ \begin{smallmatrix} i \\ kj \end{smallmatrix} \right\} \mathbf{e}^k(\mathbf{x}), \quad (31.29)$$

$$\left\{ \begin{smallmatrix} k \\ ij \end{smallmatrix} \right\} = \left\{ \begin{smallmatrix} k \\ ji \end{smallmatrix} \right\}, \quad (31.30)$$

$$\left\{ \begin{smallmatrix} m \\ ik \end{smallmatrix} \right\} = \frac{g^{jm}}{2} \left[\frac{\partial g_{ij}}{\partial x^k} + \frac{\partial g_{kj}}{\partial x^i} - \frac{\partial g_{ik}}{\partial x^j} \right]. \quad (31.31)$$

Proof: Formula 31.28 is the definition of the Christoffel symbols. We verify 31.29 next. To do so, note

$$\mathbf{e}^i(\mathbf{x}) \cdot \mathbf{e}_k(\mathbf{x}) = \delta_k^i.$$

Then from the product rule,

$$\frac{\partial \mathbf{e}^i(\mathbf{x})}{\partial x^j} \cdot \mathbf{e}_k(\mathbf{x}) + \mathbf{e}^i(\mathbf{x}) \cdot \frac{\partial \mathbf{e}_k(\mathbf{x})}{\partial x^j} = 0.$$

Now from the definition,

$$\frac{\partial e^i(x)}{\partial x^j} \cdot e_k(x) = -e^i(x) \cdot \left\{ \begin{matrix} r \\ kj \end{matrix} \right\} e_r(x) = -\left\{ \begin{matrix} r \\ kj \end{matrix} \right\} \delta_r^i = -\left\{ \begin{matrix} i \\ kj \end{matrix} \right\}.$$

But also, using the above,

$$\frac{\partial e^i(x)}{\partial x^j} = \frac{\partial e^i(x)}{\partial x^j} \cdot e_k(x) e^k(x) = -\left\{ \begin{matrix} i \\ kj \end{matrix} \right\} e^k(x).$$

This verifies 31.29. Formula 31.30 follows from 31.26 and equality of mixed partial derivatives.

It remains to show 31.31.

$$\frac{\partial g_{ij}}{\partial x^k} = \frac{\partial e_i}{\partial x^k} \cdot e_j + e_i \cdot \frac{\partial e_j}{\partial x^k} = \left\{ \begin{matrix} r \\ ik \end{matrix} \right\} e_r \cdot e_j + e_i \cdot e_r \left\{ \begin{matrix} r \\ jk \end{matrix} \right\}.$$

Therefore,

$$\frac{\partial g_{ij}}{\partial x^k} = \left\{ \begin{matrix} r \\ ik \end{matrix} \right\} g_{rj} + \left\{ \begin{matrix} r \\ jk \end{matrix} \right\} g_{ri}. \quad (31.32)$$

Switching i and k while remembering 31.30 yields

$$\frac{\partial g_{kj}}{\partial x^i} = \left\{ \begin{matrix} r \\ ik \end{matrix} \right\} g_{rj} + \left\{ \begin{matrix} r \\ ji \end{matrix} \right\} g_{rk}. \quad (31.33)$$

Now switching j and k in 31.32,

$$\frac{\partial g_{ik}}{\partial x^j} = \left\{ \begin{matrix} r \\ ij \end{matrix} \right\} g_{rk} + \left\{ \begin{matrix} r \\ jk \end{matrix} \right\} g_{ri}. \quad (31.34)$$

Adding 31.32 to 31.33 and subtracting 31.34 yields

$$\frac{\partial g_{ij}}{\partial x^k} + \frac{\partial g_{kj}}{\partial x^i} - \frac{\partial g_{ik}}{\partial x^j} = 2 \left\{ \begin{matrix} r \\ ik \end{matrix} \right\} g_{rj}.$$

Now multiplying both sides by g^{jm} and using the fact shown earlier in Theorem 31.1.6 that $g_{rj}g^{jm} = \delta_r^m$, it follows

$$2 \left\{ \begin{matrix} m \\ ik \end{matrix} \right\} = g^{jm} \left(\frac{\partial g_{ij}}{\partial x^k} + \frac{\partial g_{kj}}{\partial x^i} - \frac{\partial g_{ik}}{\partial x^j} \right)$$

which proves 31.31. ■

This is a very interesting formula because it shows the Christoffel symbols are completely determined by the metric tensor and its partial derivatives which illustrates the fundamental nature of the metric tensor. Note that the inner product is determined by this metric tensor.

31.7 Gradients and Divergence

The purpose of this section is to express the gradient and the divergence of a vector field in general curvilinear coordinates. As before, (y^1, \dots, y^n) will denote the standard coordinates with respect to the usual basis vectors. Thus

$$\mathbf{y} \equiv y^k \mathbf{i}_k, \quad \mathbf{e}_k(\mathbf{y}) = \mathbf{i}_k = \mathbf{e}^k(\mathbf{y}).$$

Let $\phi : U \rightarrow \mathbb{R}$ be a differentiable scalar function, sometimes called a “scalar field” in this subject. Write $\phi(\mathbf{x})$ to denote the value of ϕ at the point whose coordinates are \mathbf{x} . The same convention is used for a vector field. Thus $\mathbf{F}(\mathbf{x})$ is the value of a vector field at the point of U determined by the coordinates \mathbf{x} . In the standard rectangular coordinates, the gradient is well understood from earlier.

$$\nabla\phi(\mathbf{y}) = \frac{\partial\phi(\mathbf{y})}{\partial y^k} \mathbf{e}^k(\mathbf{y}) = \frac{\partial\phi(\mathbf{y})}{\partial y^k} \mathbf{i}^k.$$

However, the idea is to express the gradient in arbitrary coordinates. Therefore, using the chain rule, if the coordinates of the point of U are given as \mathbf{x} ,

$$\begin{aligned} \nabla\phi(\mathbf{x}) &= \nabla\phi(\mathbf{y}) = \frac{\partial\phi(\mathbf{x})}{\partial x^r} \frac{\partial x^r}{\partial y^k} \mathbf{e}^k(\mathbf{y}) = \\ &= \frac{\partial\phi(\mathbf{x})}{\partial x^r} \frac{\partial x^r}{\partial y^k} \frac{\partial y^k}{\partial x^s} \mathbf{e}^s(\mathbf{x}) = \frac{\partial\phi(\mathbf{x})}{\partial x^r} \delta_s^r \mathbf{e}^s(\mathbf{x}) = \frac{\partial\phi(\mathbf{x})}{\partial x^r} \mathbf{e}^r(\mathbf{x}). \end{aligned}$$

This shows the covariant components of $\nabla\phi(\mathbf{x})$ are

$$(\nabla\phi(\mathbf{x}))_r = \frac{\partial\phi(\mathbf{x})}{\partial x^r}, \quad (31.35)$$

Formally the same as in rectangular coordinates. To find the contravariant components, “raise the index” in the usual way. Thus

$$(\nabla\phi(\mathbf{x}))^r = g^{rk}(\mathbf{x}) (\nabla\phi(\mathbf{x}))_k = g^{rk}(\mathbf{x}) \frac{\partial\phi(\mathbf{x})}{\partial x^k}. \quad (31.36)$$

What about the divergence of a vector field? The divergence of a vector field \mathbf{F} defined on U is a scalar field, $\text{div}(\mathbf{F})$ which from calculus is

$$\frac{\partial F^k}{\partial y^k}(\mathbf{y}) = F_{,k}^k(\mathbf{y})$$

in terms of the usual rectangular coordinates \mathbf{y} . The reason the above equation holds in this case is that $\mathbf{e}_k(\mathbf{y})$ is a constant and so the Christoffel symbols are zero. We want an expression for the divergence in arbitrary coordinates. From Theorem 31.6.1,

$$F_{,j}^i(\mathbf{y}) = F_{,s}^r(\mathbf{x}) \frac{\partial x^s}{\partial y^j} \frac{\partial y^i}{\partial x^r}$$

From 31.27,

$$= \left(\frac{\partial F^r(\mathbf{x})}{\partial x^s} + F^k(\mathbf{x}) \left\{ \begin{matrix} r \\ ks \end{matrix} \right\}(\mathbf{x}) \right) \frac{\partial x^s}{\partial y^j} \frac{\partial y^i}{\partial x^r}.$$

Letting $j = i$ yields

$$\begin{aligned} \text{div}(\mathbf{F}) &= \left(\frac{\partial F^r(\mathbf{x})}{\partial x^s} + F^k(\mathbf{x}) \left\{ \begin{matrix} r \\ ks \end{matrix} \right\}(\mathbf{x}) \right) \frac{\partial x^s}{\partial y^i} \frac{\partial y^i}{\partial x^r} \\ &= \left(\frac{\partial F^r(\mathbf{x})}{\partial x^s} + F^k(\mathbf{x}) \left\{ \begin{matrix} r \\ ks \end{matrix} \right\}(\mathbf{x}) \right) \delta_r^s \\ &= \left(\frac{\partial F^r(\mathbf{x})}{\partial x^r} + F^k(\mathbf{x}) \left\{ \begin{matrix} r \\ kr \end{matrix} \right\}(\mathbf{x}) \right). \end{aligned} \quad (31.37)$$

$\left\{ \begin{smallmatrix} r \\ kr \end{smallmatrix} \right\}$ is simplified using the description of it in Theorem 31.6.2. Thus, from this theorem,

$$\left\{ \begin{smallmatrix} r \\ rk \end{smallmatrix} \right\} = \frac{g^{jr}}{2} \left[\frac{\partial g_{rj}}{\partial x^k} + \frac{\partial g_{kj}}{\partial x^r} - \frac{\partial g_{rk}}{\partial x^j} \right]$$

Now consider $\frac{g^{jr}}{2}$ times the last two terms in $[\cdot]$. Relabeling the indices r and j in the second term implies

$$\frac{g^{jr}}{2} \frac{\partial g_{kj}}{\partial x^r} - \frac{g^{jr}}{2} \frac{\partial g_{rk}}{\partial x^j} = \frac{g^{jr}}{2} \frac{\partial g_{kj}}{\partial x^r} - \frac{g^{rj}}{2} \frac{\partial g_{jk}}{\partial x^r} = 0.$$

Therefore,

$$\left\{ \begin{smallmatrix} r \\ rk \end{smallmatrix} \right\} = \frac{g^{jr}}{2} \frac{\partial g_{rj}}{\partial x^k}. \quad (31.38)$$

Now recall $g \equiv \det(g_{ij}) = \det(G) > 0$ from Theorem 31.1.6. Also from the formula for the inverse of a matrix and this theorem,

$$g^{jr} = A^{rj} (\det G)^{-1} = A^{jr} (\det G)^{-1}$$

where A^{rj} is the rj^{th} cofactor of the matrix (g_{ij}) . Also recall that

$$g = \sum_{r=1}^n g_{rj} A^{rj} \text{ no sum on } j.$$

Therefore, g is a function of the variables $\{g_{rj}\}$ and $\frac{\partial g}{\partial g_{rj}} = A^{rj}$. From 31.38,

$$\left\{ \begin{smallmatrix} r \\ rk \end{smallmatrix} \right\} = \frac{g^{jr}}{2} \frac{\partial g_{rj}}{\partial x^k} = \frac{1}{2g} \frac{\partial g_{rj}}{\partial x^k} A^{jr} = \frac{1}{2g} \frac{\partial g}{\partial g_{rj}} \frac{\partial g_{rj}}{\partial x^k} = \frac{1}{2g} \frac{\partial g}{\partial x^k}$$

and so from 31.37,

$$\begin{aligned} \operatorname{div}(\mathbf{F}) &= \frac{\partial F^k(\mathbf{x})}{\partial x^k} + \\ &+ F^k(\mathbf{x}) \frac{1}{2g(\mathbf{x})} \frac{\partial g(\mathbf{x})}{\partial x^k} = \frac{1}{\sqrt{g(\mathbf{x})}} \frac{\partial}{\partial x^i} \left(F^i(\mathbf{x}) \sqrt{g(\mathbf{x})} \right). \end{aligned} \quad (31.39)$$

This is the formula for the divergence of a vector field in general curvilinear coordinates. Note that it uses the contravariant components of \mathbf{F} .

The Laplacian of a scalar field is nothing more than the divergence of the gradient. In symbols, $\Delta\phi \equiv \nabla \cdot \nabla\phi$. From 31.39 and 31.36 it follows

$$\Delta\phi(\mathbf{x}) = \frac{1}{\sqrt{g(\mathbf{x})}} \frac{\partial}{\partial x^i} \left(g^{ik}(\mathbf{x}) \frac{\partial \phi(\mathbf{x})}{\partial x^k} \sqrt{g(\mathbf{x})} \right). \quad (31.40)$$

We summarize the conclusions of this section in the following theorem.

Theorem 31.7.1 *The following hold for gradient, divergence, and Laplacian in general curvilinear coordinates.*

$$(\nabla\phi(\mathbf{x}))_r = \frac{\partial \phi(\mathbf{x})}{\partial x^r}, \quad (31.41)$$

$$(\nabla\phi(\mathbf{x}))^r = g^{rk}(\mathbf{x}) \frac{\partial\phi(\mathbf{x})}{\partial x^k}, \quad (31.42)$$

$$\operatorname{div}(\mathbf{F}) = \frac{1}{\sqrt{g(\mathbf{x})}} \frac{\partial}{\partial x^i} \left(F^i(\mathbf{x}) \sqrt{g(\mathbf{x})} \right), \quad (31.43)$$

$$\Delta\phi(\mathbf{x}) = \frac{1}{\sqrt{g(\mathbf{x})}} \frac{\partial}{\partial x^i} \left(g^{ik}(\mathbf{x}) \frac{\partial\phi(\mathbf{x})}{\partial x^k} \sqrt{g(\mathbf{x})} \right). \quad (31.44)$$

Example 31.7.2 Define curvilinear coordinates as follows

$$x = r \cos \theta, y = r \sin \theta$$

Find $\nabla^2 f(r, \theta)$. That is, find the Laplacian in terms of these new variables r, θ .

First find the metric tensor. From the definition, this is

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

The contravariant components of the gradient are

$$\begin{pmatrix} 1 & 0 \\ 0 & r^{-2} \end{pmatrix} \begin{pmatrix} f_r \\ f_\theta \end{pmatrix} = \begin{pmatrix} f_r \\ \frac{1}{r^2} f_\theta \end{pmatrix}$$

Then also $\sqrt{g} = r$. Therefore, using the formula,

$$\nabla^2 f(u, v) = \frac{1}{r} \left[(rf_r)_r + \left(r \frac{1}{r^2} f_\theta \right)_\theta \right] = \frac{1}{r} (rf_r)_r + \frac{1}{r^2} f_{\theta\theta}$$

Notice how easy this is. It is anything but easy if you try to do it by brute force with none of the machinery developed here.

31.8 Exercises

1. Let $y^1 = x^1 + 2x^2, y^2 = x^2 + 3x^3, y^3 = x^1 + x^3$. Let

$$\mathbf{F}(\mathbf{x}) = x^1 \mathbf{e}_1(\mathbf{x}) + x^2 \mathbf{e}_2(\mathbf{x}) + (x^3)^2 \mathbf{e}(\mathbf{x}).$$

Find $\operatorname{div}(\mathbf{F})(\mathbf{x})$.

2. For the coordinates of the preceding problem, and ϕ a scalar field, find

$$(\nabla\phi(\mathbf{x}))^3$$

in terms of the partial derivatives of ϕ taken with respect to the variables x^i .

3. Let $y^1 = 7x^1 + 2x^2, y^2 = x^2 + 3x^3, y^3 = x^1 + x^3$. Let ϕ be a scalar field. Find $\nabla^2\phi(\mathbf{x})$.
4. Derive $\nabla^2 u$ in cylindrical coordinates, r, θ, z , where u is a scalar field on \mathbb{R}^3 .

$$x = r \cos \theta, y = r \sin \theta, z = z.$$

5. \uparrow Find all solutions to $\nabla^2 u = 0$ which depend only on r where $r \equiv \sqrt{x^2 + y^2}$.
6. Derive $\nabla^2 u$ in spherical coordinates.
7. \uparrow Let u be a scalar field on \mathbb{R}^3 . Find all solutions to $\nabla^2 u = 0$ which depend only on

$$\rho \equiv \sqrt{x^2 + y^2 + z^2}.$$

8. The temperature, u , in a solid satisfies $\nabla^2 u = 0$ after a long time. Suppose in a long pipe of inner radius 9 and outer radius 10 the exterior surface is held at 100° while the inner surface is held at 200° find the temperature in the solid part of the pipe.
9. Show velocity can be expressed as $\mathbf{v} = v_i(\mathbf{x}) \mathbf{e}^i(\mathbf{x})$, where

$$v_i(\mathbf{x}) = \frac{\partial r_i}{\partial x^j} \frac{dx^j}{dt} - r_p(\mathbf{x}) \left\{ \begin{matrix} p \\ ik \end{matrix} \right\} \frac{dx^k}{dt}$$

and $r_i(\mathbf{x})$ are the covariant components of the displacement vector,

$$\mathbf{r} = r_i(\mathbf{x}) \mathbf{e}^i(\mathbf{x}).$$

10. Find the covariant components of velocity in spherical coordinates. **Hint:** $\mathbf{v} = \frac{d\mathbf{y}}{dt}$. Now use chain rule and identify the contravariant components. Then use the technique of lowering or raising index.
11. Show that $\mathbf{v} \cdot \mathbf{w} = g_{ij}(\mathbf{x}) v^i(\mathbf{x}) v^j(\mathbf{x}) = g^{ij}(\mathbf{x}) v_i(\mathbf{x}) v_j(\mathbf{x})$.

Chapter 32

Measures and Integrals

If you want to understand a decent theory of integration, you need to do something other than the Riemann integral. In particular, if probability is of interest, you must understand the notion of measure theory and the abstract Lebesgue integral. It is also the case that these topics are much easier to follow than the extreme technicalities required for a rigorous description of the very inferior Riemann integral of a function of many variables. That is why I am placing this material in this elementary book. The rigorous description of the Riemann integral for functions of many variables is in my engineering math book and also in my earlier calculus book [21]. It is very technical and what you end up with is not nearly as good. The usual solution to this problem is to simply leave out the rigorous presentation and pretend people understand it when they don't. This is essentially what I did earlier in the book and you will see this done even in advanced calculus courses. I attempted to make the integral plausible through the use of iterated integrals. This required an emphasis on integration over very simple regions, those for which you can actually compute the integral, and it avoids the fundamental questions.

There are two chapters devoted to this material. The first is on the abstract framework for Lebesgue integration. It has a very different flavor than what you saw up till now. The second chapter considers the special case of Lebesgue integration and measure in \mathbb{R}^p . If you understand the first of these chapters, this one will seem fairly easy. I believe it is worth mastering the abstract material in order to gain a more up to date understanding of the integral. However, this is only an introduction. I have neglected all the very important material on representation theorems and functions spaces and regularity of the measures. You can see this in my on line book Calculus of Real and Complex Variables which is intended to follow this book or in Real and Abstract Analysis also on my web page. There are many standard texts which also give this material such as [20, 27].

Notation 32.0.1 *In this chapter Ω will be some nonempty set. It could be a subset of \mathbb{R}^p , the integers, part of a probability space, a part of a manifold, etc. First of all, the notation $[g < f]$ is short for $\{\omega \in \Omega : g(\omega) < f(\omega)\}$ 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 .*

Also $\mathcal{X}_E(\omega)$ is defined as

$$\mathcal{X}_E(\omega) \equiv \begin{cases} 1 & \text{if } \omega \in E \\ 0 & \text{if } \omega \notin E \end{cases}$$

This is called the indicator function of the set E because it indicates whether $\omega \in E$, 1 if ω is in E and 0 if it is not.

32.1 Countable Sets

There are different kinds of infinity. The smallest one is called \aleph_0 and is referred to as countably infinite. The theory of the Lebesgue integral lives in the land of sequences and the indices of these come from a countably infinite set. One considers countable intersections and unions. I will only include the minimum needed to understand measure and integration. A much more complete treatment is in Hewitt and Stromberg [20].

Definition 32.1.1 A set S is said to be countable if there is a function $f : \mathbb{N} \rightarrow S$ which is onto. This means you can assign a positive integer to each element of S . In other words, you could write $S = \bigcup_{k=1}^{\infty} \{s_k\}$. If X, Y are countable sets, then so is $X \times Y$.

Proof: This follows from the diagram in which $X = \{x_k\}$ and $Y = \{y_j\}$

$$\begin{array}{ccccccc} (x_1, y_1) & (x_1, y_2) & (x_1, y_3) & (x_1, y_4) & \cdots \\ (x_2, y_1) & (x_2, y_2) & (x_2, y_3) & (x_2, y_4) & \cdots \\ (x_3, y_1) & (x_3, y_2) & (x_3, y_3) & (x_3, y_4) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \end{array}$$

Now pick a route through this doubly infinite array of ordered pairs:

$$(x_1, y_1) (x_2, y_1) (x_1, y_2) (x_3, y_1) (x_2, y_2) (x_1, y_3) \cdots$$

You will see the pattern if you begin with the sequence just shown. Give an ordered pair the number which corresponds to its order in the above listing process. Thus you pick up the entire $X \times Y$, giving each ordered pair a number from \mathbb{N} . ■

Lemma 32.1.2 If A, B are countable, so is $A \cup B$. If A is countable and if \hat{A} is a subset of A , then \hat{A} is countable also.

Proof: Consider the array

$$\begin{array}{ccccccc} a_1 & a_2 & a_3 & a_4 & \cdots \\ b_1 & b_2 & b_3 & b_4 & \cdots \end{array}$$

Then list them as follows $a_1, b_1, a_2, b_2, \dots$. Give the number in this list the number which corresponds to its order in the listing process. As to the last claim, let

$$a_1 \quad a_2 \quad a_3 \quad a_4 \quad \cdots$$

be the list of things in A . Let a_{n_1} be the first in \hat{A} . If a_{n_1}, \dots, a_{n_k} have been chosen, $n_1 < n_2 < \dots < n_k$, let $a_{n_{k+1}}$ be the next in \hat{A} . Then $k \rightarrow a_{n_k}$ lists all the elements of \hat{A} . ■

Corollary 32.1.3 The rational numbers are countable.

Proof: The positive rational numbers are included in the set of numbers m/n where m, n are positive integers. Considering a doubly infinite array like the above, the element in the m^{th} row and n^{th} column being m/n , it follows that all positive rationals are listed. Letting 0 be the first in the list shows that all nonnegative rationals are countable. Then similarly, the non positive rationals are countable. It follows from the above lemma that the rationals are countable along with every subset of the rationals. ■

Corollary 32.1.4 \mathbb{Q}^p is countable and also dense in \mathbb{R}^p .

Proof: Let $\mathbf{x} \in \mathbb{R}^p$. Let $|r_i - x_i| < \delta$. Then $|\mathbf{r} - \mathbf{x}| \leq \left(\sum_{k=1}^p |r_i - x_i|^2\right)^{1/2} < \sqrt{p}\delta$. Therefore, contained in $B(\mathbf{x}, r)$ is a point of \mathbb{Q}^p if we make each $|r_i - x_i| < \delta$ where $\sqrt{p}\delta < r$. This shows \mathbb{Q}^p is dense in \mathbb{R}^p . As to it being countable, it was shown that \mathbb{Q} is countable. Suppose \mathbb{Q}^m is countable. Then by Proposition 32.1.1, $\mathbb{Q}^m \times \mathbb{Q}$ is also countable. It follows by induction that \mathbb{Q}^n is countable for any $n \in \mathbb{N}$. ■

Note this does not say that an infinite Cartesian product of \mathbb{Q} is countable. This is not even true. However, for any $p \in \mathbb{N}$, \mathbb{Q}^p is indeed countable.

Theorem 32.1.5 If U is an open set in \mathbb{R}^p , then U is the union of countably many open boxes of the form $\prod_{k=1}^p (a_k, b_k)$.

Proof: Let $\mathbf{x} \in U$. Then $B(\mathbf{x}, r) \subseteq U$. Pick $\mathbf{r} \in \mathbb{Q}^p$ such that $\mathbf{r} \in B\left(\mathbf{x}, \frac{r}{10^p}\right)$. Then let $u_i \in \mathbb{Q}$ be such that $|x_i - r_i| < u_i < \frac{r}{10^p}$ for each i . Consider $R \equiv \prod_{i=1}^p (r_i - u_i, r_i + u_i)$. Then $\mathbf{x} \in R$ and if $\mathbf{y} \in R$, Then

$$|\mathbf{x} - \mathbf{y}| = \left(\sum_{i=1}^p 2^2 u_i^2\right)^{1/2} \leq 2 \left(p \frac{r^2}{100p^2}\right)^{1/2} = \frac{2}{\sqrt{p}} \frac{r}{10} < r$$

Thus $\mathbf{x} \in R \subseteq B(\mathbf{x}, r) \subseteq U$. There are countably many such boxes $\prod_{i=1}^p (a_i, b_i)$ where each a_i, b_i are rational and this has just shown that every point of U is in one of these boxes which is also contained in U . ■

Note there are countably many of those boxes because there are countably many (a_i, b_i) for each $i \leq p$ and what is wanted is a finite Cartesian product of these.

Also of great importance is the following lemma which says that you can stuff any open set with half open boxes with no overlap at all.

Lemma 32.1.6 Every open set in \mathbb{R}^p is the countable disjoint union of half open boxes of the form

$$\prod_{i=1}^p (a_i, a_i + 2^{-k}]$$

where $a_i = l2^{-k}$ for some integers, l, k where $k \geq m$. If \mathcal{B}_m denotes this collection of half open boxes, then every box of \mathcal{B}_{m+1} is contained in a box of \mathcal{B}_m or equals a box of \mathcal{B}_m .

Proof: Let $m \in \mathbb{N}$ be given and let $k \geq m$.

$$\mathcal{C}_k = \{\text{All half open boxes } \prod_{i=1}^p (a_i, a_i + 2^{-k}] \text{ where}$$

$$a_i = l2^{-k} \text{ for some integer } l.\}$$

Thus \mathcal{C}_k consists of a countable disjoint collection of boxes whose union is \mathbb{R}^p . This is sometimes called a tiling of \mathbb{R}^p . Think of tiles on the floor of a bathroom and you will get the idea. Note that each box has diameter no larger than $2^{-k}\sqrt{p}$. This is because if we have two points,

$$\mathbf{x}, \mathbf{y} \in \prod_{i=1}^p (a_i, a_i + 2^{-k}],$$

then $|x_i - y_i| \leq 2^{-k}$. Therefore,

$$|\mathbf{x} - \mathbf{y}| \leq \left(\sum_{i=1}^p (2^{-k})^2 \right)^{1/2} = 2^{-k} \sqrt{p}.$$

Also, a box of \mathcal{C}_{k+1} is either contained in a box of \mathcal{C}_k or it has empty intersection with this box of \mathcal{C}_k .

Let U be open and let $\mathcal{B}_1 \equiv$ all sets of \mathcal{C}_1 which are contained in U . If $\mathcal{B}_1, \dots, \mathcal{B}_k$ have been chosen, $\mathcal{B}_{k+1} \equiv$ all sets of \mathcal{C}_{k+1} contained in

$$U \setminus \bigcup_{i=1}^k \mathcal{B}_i.$$

Let $\mathcal{B}_\infty = \bigcup_{i=1}^\infty \mathcal{B}_i$. I claim $\bigcup \mathcal{B}_\infty = U$. Clearly $\bigcup \mathcal{B}_\infty \subseteq U$ because every box of every \mathcal{B}_i is contained in U . If $p \in U$, let k be the smallest integer such that p is contained in a box from \mathcal{C}_k which is also a subset of U . Thus

$$p \in \bigcup \mathcal{B}_k \subseteq \bigcup \mathcal{B}_\infty.$$

Hence \mathcal{B}_∞ is the desired countable disjoint collection of half open boxes whose union is U . The last claim follows from the construction. ■

Note that there are countably many boxes in \mathcal{B}_∞ because they are disjoint, each contains an open set, and each of these open sets contains a point from the countably many \mathbb{Q}^p .

32.2 Simple Functions, σ Algebras, Measurability

The Riemann integral, was defined in terms of step functions. One of these is of the form

$$s(x) = \sum_{k=1}^n \mathcal{X}_{I_k}(x) c_k$$

where I_k is an interval. Typically we have non overlapping intervals I_k whose union is an interval $[a, b]$ and a step function has the value c_1 on I_1, c_2 on I_2 and so forth. We also know that

$$\int s(x) dx = \sum_{k=1}^n c_k \int \mathcal{X}_{I_k} dx = \sum_{k=1}^n c_k (\text{length of } I_k).$$

In defining the Riemann integral, $c_k = f(x_k)$ for some $x_k \in I_k$ and the integral exists when these approximate integrals approach a value as the lengths of all intervals converge to 0. The maximum of all lengths was the “norm of the partition”. We think of $s(x)$ as an approximation of a given function f . If f is continuous, you can verify easily, using

uniform continuity that the step function corresponding to a Riemann sum for the norm of the partition sufficiently small will be uniformly close to the function f and so the integral of the step function will be close to the integral of the function, the integral of the step function being just a Riemann sum.

A simple function looks just like a step function except the intervals I_k are replaced with sets which might not be intervals and might not even have small diameter. In the Riemann integral, this insistence of using the intervals results in not having f too far from being continuous. If you develop things in terms of simple functions, leading to the Lebesgue integral, all topological considerations are completely eliminated. This is why the Lebesgue integral is so vastly superior and so much easier to understand and use. If

$$s(\omega) = \sum_{k=1}^n \chi_{E_i}(\omega) c_k$$

then

$$\int s(\omega) d\mu \equiv \sum_{k=1}^n \mu(E_i) c_k$$

where $\mu(E_i)$ denotes the measure of E_i . This is a more general concept than length. It could refer to probability that a random vector has values in the event $E_i \subseteq \mathbb{R}^p$ for example.

We would like to be able to measure all subsets of a given set Ω but it turns out that this is usually impossible to include along with all of the following definition. This will become clear a little later in the discussion of outer measures. However, the notion of a σ algebra turns out to be the ideal thing for a theory of integration.

Definition 32.2.1 *Let Ω be a nonempty set. A σ algebra \mathcal{F} is a set whose elements are subsets of Ω which satisfies the following.*

1. *If $E_i \in \mathcal{F}$, for $i = 1, 2, \dots$, then $\cup_{i=1}^{\infty} E_i \in \mathcal{F}$.*
2. *If $E \in \mathcal{F}$, then $E^C \equiv \Omega \setminus E \in \mathcal{F}$*
3. *\emptyset, Ω are both in \mathcal{F}*

$\mu : \mathcal{F} \rightarrow [0, \infty]$ is called a measure if whenever $E_i \in \mathcal{F}$ and $E_i \cap E_j = \emptyset$ for all $i \neq j$, then

$$\mu(\cup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} \mu(E_i)$$

that sum is defined as $\sup_n \sum_{i=1}^n \mu(E_i)$. It could be a real number or $+\infty$. Such a pair (Ω, \mathcal{F}) is called a measurable space. If you add in μ , written as $(\Omega, \mathcal{F}, \mu)$, it is called a measure space.

Of course our main interest is where Ω is a nonempty subset of \mathbb{R} or \mathbb{R}^p and the measure μ is something to do with length or p dimensional volume, returning the length for an interval or volume of a p dimensional box, but it is no more trouble to present this in the generality just described and such a generalization is essential to understand if you want to study mathematical statistics or probability. Surely the study of the integral should lead somewhere.

Observation 32.2.2 If (Ω, \mathcal{F}) is a measurable space and $E_i \in \mathcal{F}$, then $\cap_{i=1}^{\infty} E_i \in \mathcal{F}$. This is because $E_i \in \mathcal{F}$ and by DeMorgan's laws,

$$\cap_{i=1}^{\infty} E_i = \left(\cup_{i=1}^{\infty} E_i^C \right)^C \in \mathcal{F} \text{ since each } E_i^C \in \mathcal{F}$$

Measures have the following fundamental property.

Lemma 32.2.3 If μ is a measure and $F_i \in \mathcal{F}$, then $\mu(\cup_{i=1}^{\infty} F_i) \leq \sum_{i=1}^{\infty} \mu(F_i)$. Also if $F_n \in \mathcal{F}$ and $F_n \subseteq F_{n+1}$ for all n , then if $F = \cup_n F_n$,

$$\mu(F) = \lim_{n \rightarrow \infty} \mu(F_n)$$

Symbolically, if $F_n \uparrow F$, then $\mu(F_n) \uparrow \mu(F)$. If $F_n \supseteq F_{n+1}$ for all n , then if $\mu(F_1) < \infty$ and $F = \cap_n F_n$, then

$$\mu(F) = \lim_{n \rightarrow \infty} \mu(F_n)$$

Symbolically, if $\mu(F_1) < \infty$ and $F_n \downarrow F$, then $\mu(F_n) \downarrow \mu(F)$.

Proof: Let $G_1 = F_1$ and if G_1, \dots, 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

$$\mu(G_{n+1}) + \mu(F_{n+1} \cap (\cup_{i=1}^n G_i)) = \mu(F_{n+1})$$

and so $\mu(G_n) \leq \mu(F_n)$. Therefore,

$$\mu(\cup_{i=1}^{\infty} G_i) = \mu(\cup_{i=1}^{\infty} F_i) = \sum_i \mu(G_i) \leq \sum_i \mu(F_i).$$

Now consider the increasing sequence of $F_n \in \mathcal{F}$. If $F \subseteq G$ and these are sets of \mathcal{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

$$\begin{aligned} \mu(F) &= \sum_{i=1}^{\infty} \mu(F_{i+1}) - \mu(F_i) + \mu(F_1) \\ &= \lim_{n \rightarrow \infty} \sum_{i=1}^n \mu(F_{i+1}) - \mu(F_i) + \mu(F_1) = \lim_{n \rightarrow \infty} \mu(F_{n+1}) \end{aligned}$$

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,

$$\mu(F_1) - \mu(F) = \mu(F_1 \setminus F) = \lim_{n \rightarrow \infty} \mu(F_1 \setminus F_n) = \lim_{n \rightarrow \infty} (\mu(F_1) - \mu(F_n))$$

This is justified because $\mu(F_1 \setminus F_n) + \mu(F_n) = \mu(F_1)$ and all numbers are finite by assumption. Hence

$$\mu(F) = \lim_{n \rightarrow \infty} \mu(F_n). \quad \blacksquare$$

Next is a discussion of the notion of a measurable function.

Notation 32.2.4 In whatever context, $f^{-1}(S) \equiv \{\omega : f(\omega) \in S\}$. It is called the inverse image of S and everything in the theory of the Lebesgue integral is formulated in terms of inverse images. For a real valued f , $f^{-1}(\lambda, \infty)$ may be written as $[f > \lambda]$.

Lemma 32.2.5 Let $f : \Omega \rightarrow (-\infty, \infty]$ where \mathcal{F} is a σ algebra of subsets of Ω . The following are equivalent.

$$\begin{aligned} f^{-1}((d, \infty]) &\in \mathcal{F} \text{ for all finite } d, \\ f^{-1}((-\infty, d)) &\in \mathcal{F} \text{ for all finite } d, \\ f^{-1}([d, \infty]) &\in \mathcal{F} \text{ for all finite } d, \\ f^{-1}((-\infty, d]) &\in \mathcal{F} \text{ for all finite } d, \\ f^{-1}((a, b)) &\in \mathcal{F} \text{ for all } a < b, -\infty < a < b < \infty. \end{aligned}$$

Definition 32.2.6 Any of these equivalent conditions in the above lemma is what is meant when we say that f is measurable.

Proof of the lemma: First note that the first and the third are equivalent. To see this, observe

$$f^{-1}([d, \infty]) = \bigcap_{n=1}^{\infty} f^{-1}((d - 1/n, \infty]),$$

and so if the first condition holds, then so does the third.

$$f^{-1}((d, \infty]) = \bigcup_{n=1}^{\infty} f^{-1}([d + 1/n, \infty]),$$

and so if the third condition holds, so does the first.

Similarly, the second and fourth conditions are equivalent. Now

$$f^{-1}((-\infty, d]) = (f^{-1}((d, \infty]))^C$$

so the first and fourth conditions are equivalent. Thus the first four conditions are equivalent and if any of them hold, then for $-\infty < a < b < \infty$,

$$f^{-1}((a, b)) = f^{-1}((-\infty, b)) \cap f^{-1}((a, \infty]) \in \mathcal{F}.$$

Finally, if the last condition holds,

$$f^{-1}([d, \infty]) = \left(\bigcup_{k=1}^{\infty} f^{-1}((-k + d, d)) \right)^C \in \mathcal{F}$$

and so the third condition holds. Therefore, all five conditions are equivalent. \blacksquare

From this, it is easy to verify that pointwise limits of a sequence of measurable functions are measurable.

Corollary 32.2.7 *If $f_n(\omega) \rightarrow f(\omega)$ where all functions have values in $(-\infty, \infty]$, then if each f_n is measurable, so is f .*

Proof: Note the following:

$$f^{-1}\left(\left(b + \frac{1}{l}, \infty\right]\right) = \bigcup_{k=1}^{\infty} \bigcap_{n \geq k} f_n^{-1}\left(\left(b + \frac{1}{l}, \infty\right]\right) \subseteq f^{-1}\left(\left[b + \frac{1}{l}, \infty\right]\right)$$

This follows from the definition of the limit. Therefore,

$$\begin{aligned} f^{-1}((b, \infty]) &= \bigcup_{l=1}^{\infty} f^{-1}\left(\left(b + \frac{1}{l}, \infty\right]\right) = \bigcup_{l=1}^{\infty} \bigcup_{k=1}^{\infty} \bigcap_{n \geq k} f_n^{-1}\left(\left(b + \frac{1}{l}, \infty\right]\right) \\ &\subseteq \bigcup_{l=1}^{\infty} f^{-1}\left(\left[b + \frac{1}{l}, \infty\right]\right) = f^{-1}((b, \infty]) \end{aligned}$$

The messy term on the middle is measurable because it consists of countable unions and intersections of measurable sets. It equals $f^{-1}((b, \infty])$ and so this last set is also measurable. By Lemma 32.2.5, f is measurable. ■

A convenient way to check measurability is in terms of limits of simple functions.

Theorem 32.2.8 *Let $f \geq 0$ be measurable. Then there exists a sequence of nonnegative simple functions $\{s_n\}$ satisfying*

$$0 \leq s_n(\omega) \tag{32.1}$$

$$\cdots s_n(\omega) \leq s_{n+1}(\omega) \cdots$$

$$f(\omega) = \lim_{n \rightarrow \infty} s_n(\omega) \text{ for all } \omega \in \Omega. \tag{32.2}$$

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

$$t_n(\omega) = \sum_{k=0}^{2^n} \frac{k}{n} \mathcal{X}_{f^{-1}([\frac{k}{n}, \frac{k+1}{n}))}(\omega) + 2^n \mathcal{X}_I(\omega).$$

Then $t_n(\omega) \leq f(\omega)$ for all ω and $\lim_{n \rightarrow \infty} t_n(\omega) = f(\omega)$ 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

$$0 \leq f(\omega) - t_n(\omega) \leq \frac{1}{n}. \tag{32.3}$$

Thus whenever $\omega \notin I$, the above inequality will hold for all n large enough. Let

$$s_1 = t_1, s_2 = \max(t_1, t_2), s_3 = \max(t_1, t_2, t_3), \dots$$

Then the sequence $\{s_n\}$ satisfies 32.1-32.2. Also each s_n has finitely many values and is measurable. To see this, note that

$$s_n^{-1}((a, \infty]) = \bigcup_{k=1}^n t_k^{-1}((a, \infty]) \in \mathcal{F}$$

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, 32.3 holds for all ω . Thus the convergence is uniform.

Now consider the converse assertion. Why is f measurable if it is the pointwise limit of an increasing sequence simple functions?

$$f^{-1}((a, \infty]) = \cup_{n=1}^{\infty} s_n^{-1}((a, \infty])$$

because $\omega \in f^{-1}((a, \infty])$ if and only if $\omega \in s_n^{-1}((a, \infty])$ for all n sufficiently large. ■

Observation 32.2.9 *If $f : \Omega \rightarrow \mathbb{R}$ then the above definition of measurability holds with no change. In this case, f never achieves the value ∞ . This is actually the case of most interest.*

Corollary 32.2.10 *If $f : \Omega \rightarrow (-\infty, \infty)$ is measurable, then there exists a sequence of simple functions $\{s_n(\omega)\}$ such that $|s_n(\omega)| \leq |f(\omega)|$ and $s_n(\omega) \rightarrow f(\omega)$.*

Proof: Let $f_+(\omega) \equiv \frac{|f(\omega)|+f(\omega)}{2}$, $f_-(\omega) \equiv \frac{|f(\omega)|-f(\omega)}{2}$. Thus $f = f_+ - f_-$ and $|f| = f_+ + f_-$. Also $f = f_+$ when $f \geq 0$ and $f = -f_-$ when $f \leq 0$. Both f_+, f_- are measurable functions. Indeed, if $a \geq 0$, $f_+^{-1}((a, \infty)) = f^{-1}((a, \infty)) \in \mathcal{F}$. If $a < 0$ then $f_+^{-1}((a, \infty)) = \Omega$. Similar considerations hold for f_- . Now let $s_n^+(\omega) \uparrow f_+(\omega)$, $s_n^-(\omega) \uparrow f_-(\omega)$ meaning these are simple functions converging respectively to f_+ and f_- which are both increasing in n and nonnegative. Thus if $s_n(\omega) \equiv s_n^+(\omega) - s_n^-(\omega)$, this converges to $f_+(\omega) - f_-(\omega)$. Also

$$|s_n(\omega)| = s_n^+(\omega) + s_n^-(\omega) \leq f_+(\omega) + f_-(\omega) = |f(\omega)| \quad \blacksquare$$

Proposition 32.2.11 *Let $f_i : \Omega \rightarrow \mathbb{R}$ be measurable, (Ω, \mathcal{F}) a measurable space, and let $g : \mathbb{R}^p \rightarrow \mathbb{R}$ be continuous. If $\mathbf{f}(\omega) = (f_1(\omega) \cdots f_p(\omega))^T$, then $g \circ \mathbf{f}$ is measurable.*

Proof: From Corollary 32.2.10 above, there are

$$s_n^i(\omega),$$

simple functions $\lim_{n \rightarrow \infty} s_n^i(\omega) = f_i(\omega)$ such that $|s_n^i(\omega)| \leq |f_i(\omega)|$. Let

$$\mathbf{s}_n(\omega) \equiv (s_n^1(\omega) \cdots s_n^p(\omega))^T$$

thus, by continuity, $g(\mathbf{s}_n(\omega)) \rightarrow g(\mathbf{f}(\omega))$ for each ω . It remains to verify that $g \circ \mathbf{s}_n$ is measurable. $g^{-1}((a, \infty))$ is an open subset of \mathbb{R}^p and so by Theorem 32.1.5, it is a countable union of open boxes of the form $R_k = \prod_{i=1}^p (u_i^k, v_i^k)$. Thus

$$g \circ \mathbf{s}_n^{-1}((a, \infty)) = \{\omega : \mathbf{s}_n(\omega) \in \cup_{k=1}^{\infty} R_k\} = \cup_k \mathbf{s}_n^{-1}(R_k) = \cup_{k=1}^{\infty} \cap_{i=1}^p (s_n^i)^{-1}(u_i^k, v_i^k).$$

Now $(s_n^i)^{-1}(u_i^k, v_i^k)$ consists of a finite union of measurable sets because s_n^i has finitely many values on measurable sets, and so it is measurable. Hence $g \circ \mathbf{s}_n$ is measurable and so it follows from Corollary 32.2.7, $g \circ \mathbf{f}$ is measurable because it is the limit of functions which are. ■

Note how this shows as a very special case that linear combinations of measurable real valued functions are measurable because you could take $g(x, y) \equiv ax + by$ and then if you have two measurable functions f_1, f_2 , it follows that $af_1 + bf_2$ is measurable.

Definition 32.2.12 $f : \Omega \rightarrow \mathbb{R}^p$ is measurable means that each component function is real valued and measurable.

Proposition 32.2.13 $f : \Omega \rightarrow \mathbb{R}^p$ is measurable where (Ω, \mathcal{F}) is a measurable space if and only if $f^{-1}(\text{open set}) \in \mathcal{F}$.

Proof: If each component function is measurable, then

$$f^{-1}\left(\prod_{k=1}^p (a_k, b_k)\right) = \cap_{k=1}^p f_k^{-1}(a_k, b_k) \in \mathcal{F}$$

By Theorem 32.1.5, every open set U is a countable union of open rectangles $\{R_i\}$ so $f^{-1}(U) = \cup_i f^{-1}(R_i) \in \mathcal{F}$. Conversely, if $f^{-1}(\text{open set})$ is always measurable, then

$$f_k^{-1}(a, b) = f^{-1}(\mathbb{R} \times \cdots \times (a, b) \times \cdots \times \mathbb{R})$$

is measurable so the component functions are measurable. ■

32.3 Measures and Outer Measures

There is also something called an outer measure which is defined on the set of all subsets.

Definition 32.3.1 Let Ω be a nonempty set and let $\lambda : \mathcal{P}(\Omega) \rightarrow [0, \infty)$ 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.

It is just like a measure except you only get 3. and you do not know that you get equality if the E_i are disjoint. Every measure determines an outer measure. For example, suppose that μ is a measure on \mathcal{F} a σ algebra of subsets of Ω . Then define

$$\hat{\mu}(S) \equiv \inf \{ \mu(E) : E \supseteq S, E \in \mathcal{F} \}$$

This is easily seen to be an outer measure. Also, we have the following Proposition.

Proposition 32.3.2 Let μ be a measure as just described. Then $\hat{\mu}$ as defined above, is an outer measure and also, if $E \in \mathcal{F}$, then $\hat{\mu}(E) = \mu(E)$.

Proof: The first two properties of an outer measure are obvious. What of the third? If any $\hat{\mu}(E_i) = \infty$, then there is nothing to show so suppose each of these is finite. Let $F_i \supseteq E_i$ such that $F_i \in \mathcal{F}$ and $\hat{\mu}(E_i) + \frac{\varepsilon}{2^i} > \mu(F_i)$. Then

$$\begin{aligned} \hat{\mu}(\cup_{i=1}^{\infty} E_i) &\leq \mu(\cup_{i=1}^{\infty} F_i) \leq \sum_{i=1}^{\infty} \mu(F_i) \\ &< \sum_{i=1}^{\infty} \left(\hat{\mu}(E_i) + \frac{\varepsilon}{2^i} \right) = \sum_{i=1}^{\infty} \hat{\mu}(E_i) + \varepsilon \end{aligned}$$

Since ε is arbitrary, this establishes the third condition. Finally, if $E \in \mathcal{F}$, then by definition, $\hat{\mu}(E) \leq \mu(E)$ because $E \supseteq E$. Also, $\mu(E) \leq \mu(F)$ for all $F \in \mathcal{F}$ such that $F \supseteq E$. It follows that $\mu(E)$ is a lower bound of all such $\mu(F)$ and so $\hat{\mu}(E) \geq \mu(E)$. ■

32.4 Measures from Outer Measures

Earlier in Theorem 33.1.1 an outer measure on $\mathcal{P}(\mathbb{R})$ was constructed. This can be used to obtain a measure defined on \mathbb{R} . However, the procedure for doing so is a special case of a general approach due to Caratheodory about 1918.

Definition 32.4.1 *Let Ω be a nonempty set and let $\mu : \mathcal{P}(\Omega) \rightarrow [0, \infty]$ be an outer measure. For $E \subseteq \Omega$, E is μ measurable if for all $S \subseteq \Omega$,*

$$\mu(S) = \mu(S \setminus E) + \mu(S \cap E). \quad (32.4)$$

To help in remembering 32.4, 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 32.4. In the Bible, there are several incidents recorded in which a process of division resulted in more stuff than was originally present.¹ We don't want this. Measurable sets are exactly those which are incapable of such a miracle. You might think of the measurable sets as the non-miraculous sets. The idea is to show that they form a σ algebra on which the outer measure μ is a measure.

First here is a definition and a lemma.

Definition 32.4.2 $(\mu|_S)(A) \equiv \mu(S \cap A)$ for all $A \subseteq \Omega$. Thus $\mu|_S$ 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 32.4.3 *If A is μ measurable, then A is $\mu|_S$ measurable.*

Proof: Suppose A is μ measurable. It is desired to show that for all $T \subseteq \Omega$,

$$(\mu|_S)(T) = (\mu|_S)(T \cap A) + (\mu|_S)(T \setminus A).$$

Thus it is desired to show

$$\mu(S \cap T) = \mu(T \cap A \cap S) + \mu(T \cap S \cap A^C). \quad (32.5)$$

But 32.5 holds because A is μ measurable. Apply Definition 32.4.1 to $S \cap T$ instead of S .

■

If A is $\mu|_S$ measurable, it does not follow that A is μ measurable. Indeed, if you believe in the existence of non measurable sets, you could let $A = S$ for such a μ non measurable set and verify that S is $\mu|_S$ 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 volume 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 32.4.4 Let Ω be a set and let μ be an outer measure on $\mathcal{P}(\Omega)$. The collection of μ measurable sets \mathcal{S} , forms a σ algebra and

$$\text{If } F_i \in \mathcal{S}, F_i \cap F_j = \emptyset, \text{ then } \mu(\cup_{i=1}^{\infty} F_i) = \sum_{i=1}^{\infty} \mu(F_i). \quad (32.6)$$

If $\cdots F_n \subseteq F_{n+1} \subseteq \cdots$, then if $F = \cup_{n=1}^{\infty} F_n$ and $F_n \in \mathcal{S}$, it follows that

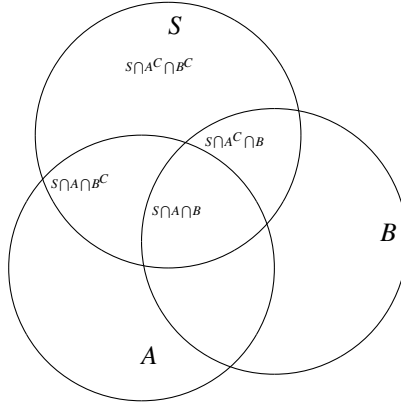
$$\mu(F) = \lim_{n \rightarrow \infty} \mu(F_n). \quad (32.7)$$

If $\cdots F_n \supseteq F_{n+1} \supseteq \cdots$, and if $F = \cap_{n=1}^{\infty} F_n$ for $F_n \in \mathcal{S}$ then if $\mu(F_1) < \infty$,

$$\mu(F) = \lim_{n \rightarrow \infty} \mu(F_n). \quad (32.8)$$

This measure space is also complete which means that if $\mu(F) = 0$ for some $F \in \mathcal{S}$ then if $G \subseteq F$, it follows $G \in \mathcal{S}$ also.

Proof: First note that \emptyset and Ω are obviously in \mathcal{S} . Now suppose $A, B \in \mathcal{S}$. I will show $A \setminus B \equiv A \cap B^C$ is in \mathcal{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

$$(S \cap A^C \cap B^C) \cup (S \cap A \cap B) \cup (S \cap A^C \cap B)$$

Therefore,

$$\begin{aligned} \mu(S) &\leq \mu(S \setminus (A \setminus B)) + \mu(S \cap (A \setminus B)) \\ &\leq \mu(S \cap A^C \cap B^C) + \mu(S \cap A \cap B) + \mu(S \cap A^C \cap B) + \mu(S \cap A \cap B) \\ &= \mu(S \cap A^C \cap B^C) + \mu(S \cap A \cap B) + \mu(S \cap A^C \cap B) + \mu(S \cap A \cap B^C) \\ &= \mu(S \cap A^C \cap B^C) + \mu(S \cap A \cap B^C) + \mu(S \cap A \cap B) + \mu(S \cap A^C \cap B) \\ &= \mu(S \cap B^C) + \mu(S \cap B) = \mu(S) \end{aligned}$$

and so this shows that $A \setminus B \in \mathcal{S}$ whenever $A, B \in \mathcal{S}$.

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 $A_1, \dots, A_n \in \mathcal{S}$, 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 $A_i \in \mathcal{S}$,

$$\mu(\cup_{i=1}^n A_i) = \sum_{i=1}^n \mu(A_i). \quad (32.9)$$

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 $n \rightarrow \infty$ and conclude,

$$\sum_{i=1}^\infty \mu(A_i) = \mu(A)$$

which establishes 32.6.

Consider part 32.7. 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 32.9

$$\mu(F_{k+1} \setminus F_k) + \mu(F_k) = \mu(F_{k+1})$$

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 32.6 just shown that

$$\begin{aligned} \mu(F) &= \sum_{k=0}^\infty \mu(F_{k+1} \setminus F_k) = \lim_{n \rightarrow \infty} \sum_{k=0}^n \mu(F_{k+1} \setminus F_k) \\ &= \lim_{n \rightarrow \infty} \sum_{k=0}^n \mu(F_{k+1}) - \mu(F_k) = \lim_{n \rightarrow \infty} \mu(F_{n+1}). \end{aligned}$$

In order to establish 32.8, let the F_n be as given there. Then, since $(F_1 \setminus F_n)$ increases to $(F_1 \setminus F)$, 32.7 implies

$$\lim_{n \rightarrow \infty} (\mu(F_1) - \mu(F_n)) = \mu(F_1 \setminus F).$$

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

$$\lim_{n \rightarrow \infty} (\mu(F_1) - \mu(F_n)) = \mu(F_1 \setminus F) \geq \mu(F_1) - \mu(F)$$

which implies

$$\lim_{n \rightarrow \infty} \mu(F_n) \leq \mu(F).$$

But since $F \subseteq F_n$,

$$\mu(F) \leq \lim_{n \rightarrow \infty} \mu(F_n)$$

and this establishes 32.8. 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

$$\begin{aligned} \mu(S) &= \mu(S \cap B_n) + \mu(S \setminus B_n) \\ &= (\mu \lfloor S)(B_n) + (\mu \lfloor S)(B_n^C). \end{aligned} \quad (32.10)$$

By Lemma 32.4.3 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 32.8 and 32.7 to the outer measure $\mu \lfloor S$ in 32.10 and let $n \rightarrow \infty$. Thus

$$B_n \uparrow A, \quad B_n^C \downarrow A^C$$

and this yields $\mu(S) = (\mu \lfloor S)(A) + (\mu \lfloor S)(A^C) = \mu(S \cap A) + \mu(S \setminus A)$.

Therefore $A \in \mathcal{S}$ and this proves Parts 32.6, 32.7, and 32.8.

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, $\mu(F \setminus G) \leq \mu(F) = 0$. ■

The measure m which results from the outer measure of Theorem 33.1.1 is called Lebesgue measure. The following is a general result about completion of a measure space. This is coming up, but first is another general result about completion of a measure space.

Proposition 32.4.5 *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Also let $\hat{\mu}$ be the outer measure defined by*

$$\hat{\mu}(F) \equiv \inf \{ \mu(E) : E \supseteq F \text{ and } E \in \mathcal{F} \}$$

Then $\hat{\mu}$ is an outer measure which is a measure on $\hat{\mathcal{F}}$, the set of $\hat{\mu}$ measurable sets. Also $\hat{\mu}(E) = \mu(E)$ for $E \in \mathcal{F}$ and $\mathcal{F} \subseteq \hat{\mathcal{F}}$. If $(\Omega, \mathcal{F}, \mu)$ is already complete, then no new sets are obtained from this process and $\mathcal{F} = \hat{\mathcal{F}}$.

Proof: The first part of this follows from Proposition 32.3.2. It only remains to verify that $\mathcal{F} \subseteq \hat{\mathcal{F}}$. Let S be a set and let $E \in \mathcal{F}$, $E_S \supseteq S$, $E_S \in \mathcal{F}$. Then

$$\mu(E_S) = \mu(E_S \setminus E) + \mu(E_S \cap E)$$

due to the fact that μ is a measure. As usual, if $\hat{\mu}(S) = \infty$, it is obvious that $\hat{\mu}(S) \geq \hat{\mu}(S \setminus E) + \hat{\mu}(S \cap E)$. Therefore, assume this is not ∞ . Then let $\hat{\mu}(S) > \mu(E_S) - \varepsilon$. Then from the above,

$$\varepsilon + \hat{\mu}(S) \geq \mu(E_S \setminus E) + \mu(E_S \cap E) \geq \mu(S \setminus E) + \mu(S \cap E)$$

Since ε is arbitrary, this shows that $E \in \hat{\mathcal{F}}$. Thus $\mathcal{F} \subseteq \hat{\mathcal{F}}$.

Why are these two σ algebras equal if $(\Omega, \mathcal{F}, \mu)$ is complete? Suppose now that $(\Omega, \mathcal{F}, \mu)$ is complete. Let $F \in \hat{\mathcal{F}}$. Then there exists $E \supseteq F$ such that $\mu(E) = \hat{\mu}(F)$. This is obvious if $\hat{\mu}(F) = \infty$. Otherwise, let $E_n \supseteq F$, $\hat{\mu}(F) + \frac{1}{n} > \mu(E_n)$. Just let $E = \bigcap_n E_n$. Now $\hat{\mu}(E \setminus F) = 0$. Now also, there exists a set of \mathcal{F} called W such that $\mu(W) = 0$ and $W \supseteq E \setminus F$. Thus $E \setminus F \subseteq W$, a set of measure zero. Hence by completeness of $(\Omega, \mathcal{F}, \mu)$, it must be the case that $E \setminus F = E \cap F^C = G \in \mathcal{F}$. Then taking complements of both sides, $E^C \cup F = G^C \in \mathcal{F}$. Now take intersections with E . $F \in E \cap G^C \in \mathcal{F}$. ■

32.5 Riemann Integrals for Decreasing Functions

A decreasing function is always Riemann integrable. This is discussed in Proposition 7.3.8. I will define the Lebesgue integral for a nonnegative function in terms of an improper Riemann integral which involves a decreasing function.

Definition 32.5.1 Let $f : [a, b] \rightarrow [0, \infty]$ be decreasing. Define

$$\int_a^b f(\lambda) d\lambda \equiv \lim_{M \rightarrow \infty} \int_a^b M \wedge f(\lambda) d\lambda = \sup_M \int_a^b M \wedge f(\lambda) d\lambda$$

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 $[0, \infty)$,

$$\int_0^\infty f d\lambda \equiv \lim_{R \rightarrow \infty} \int_0^R f d\lambda = \sup_{R > 1} \int_0^R f d\lambda = \sup_R \sup_{M > 0} \int_0^R f \wedge M d\lambda$$

Now here is an obvious property.

Lemma 32.5.2 Let f be a decreasing nonnegative function defined on an interval $[a, b]$. Then if $[a, b] = \bigcup_{k=1}^m I_k$ where $I_k \equiv [a_k, b_k]$ and the intervals I_k are non overlapping, it follows

$$\int_a^b f d\lambda = \sum_{k=1}^m \int_{a_k}^{b_k} f d\lambda.$$

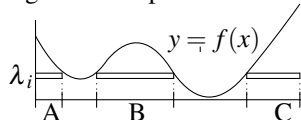
Proof: This follows from the computation,

$$\begin{aligned}\int_a^b f d\lambda &\equiv \lim_{M \rightarrow \infty} \int_a^b f \wedge M d\lambda \\ &= \lim_{M \rightarrow \infty} \sum_{k=1}^m \int_{a_k}^{b_k} f \wedge M d\lambda = \sum_{k=1}^m \int_{a_k}^{b_k} f d\lambda\end{aligned}$$

Note both sides could equal $+\infty$. ■

32.6 Lebesgue Integrals of Nonnegative Functions

Here is the definition of the Lebesgue integral of a function which is measurable and has values in $[0, \infty]$. The idea is motivated by the following picture in which $f^{-1}(\lambda_i, \infty)$ is $A \cup B \cup C$ and we take the measure of this set, multiply by $\lambda_i - \lambda_{i-1}$ and do this for each λ_i in an increasing sequence of points, $\lambda_0 \equiv 0$. Then we add the “areas” of the little horizontal “rectangles” in order to approximate the “area” under the curve. The difference here is that the “rectangles” in the sum are horizontal whereas with the Riemann integral, they are vertical. Note how it is important to be able to measure $f^{-1}(\lambda, \infty) \equiv \{x : f(x) > \lambda\} \equiv [f > \lambda]$ which is what it means for f to be measurable. Also note that, in spite of the picture, in general we don’t know a good description of this set other than that it is measurable.



Definition 32.6.1 Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and suppose $f : \Omega \rightarrow [0, \infty]$ is measurable. Then define

$$\int f d\mu \equiv \int_0^\infty \mu([f > \lambda]) d\lambda = \int_0^\infty \mu(f^{-1}(\lambda, \infty)) d\lambda$$

which makes sense because $\lambda \rightarrow \mu([f > \lambda])$ is nonnegative and decreasing. On the right you have an improper Riemann integral like what was discussed above.

Note that if $f \leq g$, then $\int f d\mu \leq \int g d\mu$ because $\mu([f > \lambda]) \leq \mu([g > \lambda])$. Next I point out that the integral is a limit of lower sums.

Lemma 32.6.2 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

$$\lim_{R \rightarrow \infty} m(h, R) = \infty$$

and so

$$\int f d\mu \equiv \int_0^\infty \mu([f > \lambda]) d\lambda = \sup_M \sup_R \int_0^R \mu([f > \lambda]) \wedge M d\lambda =$$

$$\begin{aligned}
& \sup_M \sup_{R>0} \sup_{h>0} \sum_{k=1}^{m(h,R)} (\mu([f > kh]) \wedge M) h + (\mu([f > R]) \wedge M) (R - hm(h,R)) \quad (32.11) \\
&= \sup_M \sup_{R>0} \sup_{h>0} \sum_{k=1}^{m(h,R)} (\mu([f > kh]) \wedge M) h
\end{aligned}$$

because the sum in 32.11 is just a lower sum for the integral $\int_0^R \mu([f > \lambda]) \wedge M d\lambda$, these lower sums are increasing, and the last term is smaller than Mh . Hence, switching the order of the sups, this equals

$$\begin{aligned}
& \sup_{R>0} \sup_{h>0} \sup_M \sum_{k=1}^{m(h,R)} (\mu([f > kh]) \wedge M) h = \sup_{R>0} \sup_{h>0} \lim_{M \rightarrow \infty} \sum_{k=1}^{m(h,R)} (\mu([f > kh]) \wedge M) h \\
&= \sup_{h>0} \sup_R \sum_{k=1}^{m(R,h)} (\mu([f > kh])) h = \sup_{h>0} \sum_{k=1}^{\infty} (\mu([f > kh])) h. \blacksquare
\end{aligned}$$

32.7 Nonnegative Simple Functions

To begin with, here is a useful lemma.

Lemma 32.7.1 *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,

$$\begin{aligned}
\int_0^{\infty} f(\lambda) d\lambda &= \lim_{R \rightarrow \infty} \int_0^R f(\lambda) d\lambda = \sup_{R>1} \int_0^R f(\lambda) d\lambda \\
&= \sup_{R>1} \sup_M \int_0^R f(\lambda) \wedge M d\lambda \\
&= \sup_M \sup_{R>1} \int_0^R f(\lambda) \wedge M d\lambda \\
&= \sup_M \sup_{R>1} \int_0^a f(\lambda) \wedge M d\lambda \\
&= \sup_M \int_0^a f(\lambda) \wedge M d\lambda \equiv \int_0^a f(\lambda) d\lambda. \blacksquare
\end{aligned}$$

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 32.7.2 Let $s(\omega) = \sum_{i=1}^p a_i \chi_{E_i}(\omega)$ be a nonnegative simple function where the E_i are distinct but the a_i might not be. Then

$$\int s d\mu = \sum_{i=1}^p a_i \mu(E_i). \quad (32.12)$$

Proof: Without loss of generality, assume $0 \equiv a_0 < a_1 \leq a_2 \leq \dots \leq a_p$ and that $\mu(E_i) < \infty, i > 0$. Here is why. If $\mu(E_i) = \infty$, then the left side would be

$$\begin{aligned} \int_0^{a_p} \mu([s > \lambda]) d\lambda &\geq \int_0^{a_i} \mu([s > \lambda]) d\lambda \\ &= \sup_M \int_0^{a_i} \mu([s > \lambda]) \wedge M d\lambda \\ &\geq \sup_M M a_i = \infty \end{aligned}$$

and so both sides are equal to ∞ . Thus it can be assumed that for each $i, \mu(E_i) < \infty$. Then it follows from Lemma 32.7.1 and Lemma 32.5.2,

$$\begin{aligned} \int_0^\infty \mu([s > \lambda]) d\lambda &= \int_0^{a_p} \mu([s > \lambda]) d\lambda = \sum_{k=1}^p \int_{a_{k-1}}^{a_k} \mu([s > \lambda]) d\lambda \\ &= \sum_{k=1}^p (a_k - a_{k-1}) \sum_{i=k}^p \mu(E_i) = \sum_{i=1}^p \mu(E_i) \sum_{k=1}^i (a_k - a_{k-1}) = \sum_{i=1}^p a_i \mu(E_i) \blacksquare \end{aligned}$$

Lemma 32.7.3 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

$$s(\omega) = \sum_{i=1}^n \alpha_i \chi_{A_i}(\omega), \quad t(\omega) = \sum_{j=1}^m \beta_j \chi_{B_j}(\omega)$$

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,

$$(as + bt)(\omega) = \sum_{j=1}^m \sum_{i=1}^n (a\alpha_i + b\beta_j) \chi_{A_i \cap B_j}(\omega)$$

where the sets $A_i \cap B_j$ are disjoint and measurable. By Lemma 32.7.2,

$$\begin{aligned} \int as + bt d\mu &= \sum_{j=1}^m \sum_{i=1}^n (a\alpha_i + b\beta_j) \mu(A_i \cap B_j) \\ &= \sum_{i=1}^n a \sum_{j=1}^m \alpha_i \mu(A_i \cap B_j) + b \sum_{j=1}^m \sum_{i=1}^n \beta_j \mu(A_i \cap B_j) \\ &= a \sum_{i=1}^n \alpha_i \mu(A_i) + b \sum_{j=1}^m \beta_j \mu(B_j) \\ &= a \int s d\mu + b \int t d\mu. \blacksquare \end{aligned}$$

32.8 The Monotone Convergence Theorem

The following is called the monotone convergence theorem also Beppo Levi's theorem. This theorem and related convergence theorems are the reason for using the Lebesgue integral. If $\lim_{n \rightarrow \infty} f_n(\omega) = f(\omega)$ and $f_n(\omega)$ is increasing in n , then clearly f is also measurable because

$$f^{-1}((a, \infty]) = \cup_{k=1}^{\infty} f_k^{-1}((a, \infty]) \in \mathcal{F}$$

Theorem 32.8.1 (*Monotone Convergence theorem*) Let f have values in $[0, \infty]$ and suppose $\{f_n\}$ is a sequence of nonnegative measurable functions having values in $[0, \infty]$ and satisfying

$$\begin{aligned} \lim_{n \rightarrow \infty} f_n(\omega) &= f(\omega) \text{ for each } \omega. \\ \cdots f_n(\omega) &\leq f_{n+1}(\omega) \cdots \end{aligned}$$

Then f is measurable and

$$\int f d\mu = \lim_{n \rightarrow \infty} \int f_n d\mu.$$

Proof: By Lemma 32.6.2

$$\begin{aligned} \lim_{n \rightarrow \infty} \int f_n d\mu &= \sup_n \int f_n d\mu \\ &= \sup_n \sup_{h>0} \sum_{k=1}^{\infty} \mu([f_n > kh])h = \sup_{h>0} \sup_N \sup_n \sum_{k=1}^N \mu([f_n > kh])h \\ &= \sup_{h>0} \sup_N \sum_{k=1}^N \mu([f > kh])h = \sup_{h>0} \sum_{k=1}^{\infty} \mu([f > kh])h = \int f d\mu. \blacksquare \end{aligned}$$

The next theorem, known as Fatou's lemma is another important theorem which justifies the use of the Lebesgue integral.

Theorem 32.8.2 (*Fatou's lemma*) Let f_n be a nonnegative measurable function. Let $g(\omega) = \liminf_{n \rightarrow \infty} f_n(\omega)$. Then g is measurable and

$$\int g d\mu \leq \liminf_{n \rightarrow \infty} \int f_n d\mu.$$

In other words,

$$\int \left(\liminf_{n \rightarrow \infty} f_n \right) d\mu \leq \liminf_{n \rightarrow \infty} \int f_n d\mu.$$

Proof: Let $g_n(\omega) = \inf\{f_k(\omega) : k \geq n\}$. Then

$$g_n^{-1}([a, \infty]) = \cap_{k=n}^{\infty} f_k^{-1}([a, \infty]) \in \mathcal{F}$$

Thus g_n is measurable. Now the functions g_n form an increasing sequence of nonnegative measurable functions. Thus $g^{-1}((a, \infty)) = \cup_{n=1}^{\infty} g_n^{-1}((a, \infty)) \in \mathcal{F}$ so g is measurable also. By monotone convergence theorem,

$$\int g d\mu = \lim_{n \rightarrow \infty} \int g_n d\mu \leq \liminf_{n \rightarrow \infty} \int f_n d\mu.$$

The last inequality holding because

$$\int g_n d\mu \leq \int f_n d\mu.$$

(Note that it is not known whether $\lim_{n \rightarrow \infty} \int f_n d\mu$ exists.) ■

32.9 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 32.9.1 *Let f, g be nonnegative measurable functions and let a, b be non-negative numbers. Then $af + bg$ is measurable and*

$$\int (af + bg) d\mu = a \int f d\mu + b \int g d\mu. \quad (32.13)$$

Proof: By Theorem 32.2.8 on Page 658 there exist increasing sequences of nonnegative simple functions, $s_n \rightarrow f$ and $t_n \rightarrow g$. 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 32.7.3,

$$\begin{aligned} \int (af + bg) d\mu &= \lim_{n \rightarrow \infty} \int as_n + bt_n d\mu = \lim_{n \rightarrow \infty} \left(a \int s_n d\mu + b \int t_n d\mu \right) \\ &= a \int f d\mu + b \int g d\mu. \quad \blacksquare \end{aligned}$$

32.10 Integrals of Real Valued Functions

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. The integral of real valued functions is next.

Definition 32.10.1 *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and let $f : \Omega \rightarrow \mathbb{R}$ be measurable. Then it is said to be in $L^1(\Omega, \mu)$ when*

$$\int_{\Omega} |f(\omega)| d\mu < \infty$$

Lemma 32.10.2 *If $g - h = \hat{g} - \hat{h}$ where g, \hat{g}, h, \hat{h} are measurable and nonnegative, with all integrals finite, then*

$$\int_{\Omega} g d\mu - \int_{\Omega} h d\mu = \int_{\Omega} \hat{g} d\mu - \int_{\Omega} \hat{h} d\mu$$

Proof: From Theorem 32.9.1,

$$\int \hat{g} d\mu + \int h d\mu = \int (\hat{g} + h) d\mu = \int (g + \hat{h}) d\mu = \int g d\mu + \int \hat{h} d\mu$$

and so,

$$\int \hat{g} d\mu - \int \hat{h} d\mu = \int g d\mu - \int h d\mu \quad \blacksquare$$

Definition 32.10.3 Let $f \in L^1(\Omega, \mu)$. Define $\int f d\mu \equiv \int f_+ d\mu - \int f_- d\mu$.

Proposition 32.10.4 The definition of $\int f d\mu$ is well defined and if a, b are real numbers

$$\int (af + bg) d\mu = a \int f d\mu + b \int g d\mu$$

Proof: First of all, it is well defined because f_+, f_- are both no larger than $|f|$. Therefore, $\int f_+ d\mu, \int f_- d\mu$ are both real numbers. Next, why is the integral linear. First consider the sum.

$$\int (f + g) d\mu \equiv \int (f + g)_+ d\mu - \int (f + g)_- d\mu$$

Now $(f + g)_+ - (f + g)_- = f + g = f_+ - f_- + g_+ - g_-$. By Lemma 32.10.2 and Theorem 32.9.1

$$\begin{aligned} \int (f + g) d\mu &\equiv \int (f + g)_+ d\mu - \int (f + g)_- d\mu \\ &= \int (f_+ + g_+) d\mu - \int (f_- + g_-) d\mu \\ &= \int f_+ d\mu - \int f_- d\mu + \int g_+ d\mu - \int g_- d\mu \\ &\equiv \int f d\mu + \int g d\mu \end{aligned}$$

Next note that if a is real and $a \geq 0$, $(af)_+ = af_+$, $(af)_- = af_-$ and if $a < 0$, $(af)_+ = -af_-$, $(af)_- = -af_+$. This follows from a simple computation involving the definition of f_+, f_- . Therefore, if $a < 0$,

$$\int af d\mu \equiv \int (af)_+ d\mu - \int (af)_- d\mu = \int (-a)f_- d\mu - \int (-a)f_+ d\mu$$

By Theorem 32.9.1,

$$= -a \left(\int f_- d\mu - \int f_+ d\mu \right) = a \left(\int f_+ d\mu - \int f_- d\mu \right) \equiv a \int f d\mu$$

The case where $a \geq 0$ is easier. ■

Now that we understand how to integrate real valued functions, it is time for another great convergence theorem, the dominated convergence theorem.

Theorem 32.10.5 (Dominated Convergence theorem) Let $f_n \in L^1(\Omega)$ and suppose

$$f(\omega) = \lim_{n \rightarrow \infty} f_n(\omega),$$

and there exists a measurable function g , with values in $[0, \infty]$,² such that

$$|f_n(\omega)| \leq g(\omega) \text{ and } \int g(\omega) d\mu < \infty.$$

Then $f \in L^1(\Omega)$ and

$$0 = \lim_{n \rightarrow \infty} \int |f_n - f| d\mu = \lim_{n \rightarrow \infty} \left| \int f d\mu - \int f_n d\mu \right|$$

²Note that, since g is allowed to have the value ∞ , it is not known that $g \in L^1(\Omega)$.

Proof: f is measurable by Corollary 32.2.7. Since $|f| \leq g$, it follows that

$$f \in L^1(\Omega) \text{ and } |f - f_n| \leq 2g.$$

By Fatou's lemma (Theorem 32.8.2),

$$\begin{aligned} \int 2g d\mu &\leq \liminf_{n \rightarrow \infty} \int 2g - |f - f_n| d\mu \\ &= \int 2g d\mu - \limsup_{n \rightarrow \infty} \int |f - f_n| d\mu. \end{aligned}$$

Subtracting $\int 2g d\mu$,

$$0 \leq -\limsup_{n \rightarrow \infty} \int |f - f_n| d\mu.$$

Hence

$$\begin{aligned} 0 &\geq \limsup_{n \rightarrow \infty} \left(\int |f - f_n| d\mu \right) \\ &\geq \liminf_{n \rightarrow \infty} \left(\int |f - f_n| d\mu \right) \geq \liminf_{n \rightarrow \infty} \left| \int f d\mu - \int f_n d\mu \right| \geq 0. \end{aligned}$$

This proves the theorem by Lemma 3.3.17 because the \limsup and \liminf are equal. ■

Example 32.10.6 Let $\Omega \equiv \mathbb{N}$ and let \mathcal{F} be the set of all subsets of Ω . Let $\mu(E) \equiv$ number of entries in E . Then $(\mathbb{N}, \mathcal{F}, \mu)$ is a measure space and the Lebesgue integral is summation. Thus all the convergence theorems mentioned above apply to sums.

First, why is μ a measure? If $\{E_i\}$ are disjoint, then if each is nonempty, $\cup_i E_i$ is infinite and so

$$\mu(\cup_i E_i) = \infty = \sum_{i=1}^{\infty} \mu(E_i) \geq \sum_{i=1}^{\infty} 1 = \infty$$

The alternative is that only finitely many E_i are nonempty and in this case, the assertion that $\mu(\cup_i E_i) = \sum_{i=1}^{\infty} \mu(E_i)$ is obvious. Hence μ is indeed a measure. Now let $f: \mathbb{N} \rightarrow \mathbb{R}$. It is obviously measurable because the inverse image of anything is a subset of \mathbb{N} . So if $f(n) \geq 0$ for all n , what is $\int f d\mu$?

$$f(i) = \sum_{k=1}^{\infty} f(k) \mathcal{X}_{\{k\}}(i) = \lim_{n \rightarrow \infty} \sum_{k=1}^n f(k) \mathcal{X}_{\{k\}}(i) \equiv f_n(i)$$

Now f_n is a simple function and there is exactly one thing in $\{k\}$. Therefore, $\int f_n d\mu = \sum_{k=1}^n f(k)$. Then, by the monotone convergence theorem,

$$\int f d\mu = \lim_{n \rightarrow \infty} \int f_n d\mu = \lim_{n \rightarrow \infty} \sum_{k=1}^n f(k) \equiv \sum_{k=1}^{\infty} f(k)$$

When $\sum_k |f(k)| < \infty$, one has $\int f d\mu = \sum_{k=1}^{\infty} f(k)$.

This example illustrates how the Lebesgue integral pertains to absolute summability and absolute integrability. It is not a theory which can include conditional convergence. The generalized Riemann integral, which I won't consider here can do this. However, the Lebesgue integral is very easy to use because of this restriction.

32.11 Dynkin's Lemma

Dynkin's lemma is a very useful result. It is used quite a bit in books on probability but here it is used to obtain n dimensional Lebesgue measure without any ugly technicalities.

Lemma 32.11.1 *Let \mathcal{C} be a set whose elements are σ algebras each containing some subset \mathcal{K} of the set of all subsets. Then $\cap \mathcal{C}$ is a σ algebra which contains \mathcal{K} .*

Proof: \emptyset, Ω are in $\cap \mathcal{C}$ because these are each in each σ algebra of \mathcal{C} . If $E_i \in \cap \mathcal{C}$, then if $\mathcal{F} \in \mathcal{C}$ it follows that $\cup_{i=1}^{\infty} E_i \in \mathcal{F}$ and so, since \mathcal{F} is arbitrary, this shows this union is in $\cap \mathcal{C}$. If $E \in \cap \mathcal{C}$, then $E^C \in \mathcal{F}$ for each $\mathcal{F} \in \cap \mathcal{C}$ and so, as before, $E^C \in \cap \mathcal{C}$. Thus $\cap \mathcal{C}$ is a σ algebra. ■

Definition 32.11.2 *Let Ω be a set and let \mathcal{K} be a collection of subsets of Ω . Then \mathcal{K} is called a π system if $\emptyset, \Omega \in \mathcal{K}$ and whenever $A, B \in \mathcal{K}$, it follows $A \cap B \in \mathcal{K}$. $\sigma(\mathcal{K})$ will denote the intersection of all σ algebras containing \mathcal{K} . The set of all subsets of Ω is one such σ algebra which contains \mathcal{K} . Thus $\sigma(\mathcal{K})$ is the smallest σ algebra which contains \mathcal{K} .*

The following is the fundamental lemma which shows these π systems are useful. This is due to Dynkin.

Lemma 32.11.3 *Let \mathcal{K} be a π system of subsets of Ω , a set. Also let \mathcal{G} be a collection of subsets of Ω which satisfies the following three properties.*

1. $\mathcal{K} \subseteq \mathcal{G}$
2. If $A \in \mathcal{G}$, then $A^C \in \mathcal{G}$
3. If $\{A_i\}_{i=1}^{\infty}$ is a sequence of disjoint sets from \mathcal{G} then $\cup_{i=1}^{\infty} A_i \in \mathcal{G}$.

Then $\mathcal{G} \supseteq \sigma(\mathcal{K})$, where $\sigma(\mathcal{K})$ is the smallest σ algebra which contains \mathcal{K} .

Proof: First note that if

$$\mathcal{H} \equiv \{\mathcal{G} : \text{1 - 3 all hold}\}$$

then $\cap \mathcal{H}$ yields a collection of sets which also satisfies 1 - 3. Therefore, I will assume in the argument that \mathcal{G} is the smallest collection satisfying 1 - 3. Let $A \in \mathcal{K}$ and define

$$\mathcal{G}_A \equiv \{B \in \mathcal{G} : A \cap B \in \mathcal{G}\}.$$

I want to show \mathcal{G}_A satisfies 1 - 3 because then it must equal \mathcal{G} since \mathcal{G} is the smallest collection of subsets of Ω which satisfies 1 - 3. This will give the conclusion that for $A \in \mathcal{K}$ and $B \in \mathcal{G}$, $A \cap B \in \mathcal{G}$. This information will then be used to show that if $A, B \in \mathcal{G}$ then $A \cap B \in \mathcal{G}$. From this it will follow very easily that \mathcal{G} is a σ algebra which will imply it contains $\sigma(\mathcal{K})$. Now here are the details of the argument.

Since \mathcal{K} is given to be a π system, $\mathcal{K} \subseteq \mathcal{G}_A$. Property 3 is obvious because if $\{B_i\}$ is a sequence of disjoint sets in \mathcal{G}_A , then

$$A \cap \cup_{i=1}^{\infty} B_i = \cup_{i=1}^{\infty} A \cap B_i \in \mathcal{G}$$

because $A \cap B_i \in \mathcal{G}$ and the property 3 of \mathcal{G} .

It remains to verify Property 2 so let $B \in \mathcal{G}_A$. I need to verify that $B^C \in \mathcal{G}_A$. In other words, I need to show that $A \cap B^C \in \mathcal{G}$. However,

$$A \cap B^C = (A^C \cup (A \cap B))^C \in \mathcal{G}$$

Here is why. Since $B \in \mathcal{G}_A$, $A \cap B \in \mathcal{G}$ and since $A \in \mathcal{K} \subseteq \mathcal{G}$ it follows $A^C \in \mathcal{G}$ by assumption 2. It follows from assumption 3 the union of the disjoint sets, A^C and $(A \cap B)$ is in \mathcal{G} and then from 2 the complement of their union is in \mathcal{G} . Thus \mathcal{G}_A satisfies 1 - 3 and this implies since \mathcal{G} is the smallest such, that $\mathcal{G}_A \supseteq \mathcal{G}$. However, \mathcal{G}_A is constructed as a subset of \mathcal{G} . This proves that for every $B \in \mathcal{G}$ and $A \in \mathcal{K}$, $A \cap B \in \mathcal{G}$. Now pick $B \in \mathcal{G}$ and consider

$$\mathcal{G}_B \equiv \{A \in \mathcal{G} : A \cap B \in \mathcal{G}\}.$$

I just proved $\mathcal{K} \subseteq \mathcal{G}_B$. The other arguments are identical to show \mathcal{G}_B satisfies 1 - 3 and is therefore equal to \mathcal{G} . This shows that whenever $A, B \in \mathcal{G}$ it follows $A \cap B \in \mathcal{G}$.

This implies \mathcal{G} is a σ algebra. To show this, all that is left is to verify \mathcal{G} is closed under countable unions because then it follows \mathcal{G} is a σ algebra. Let $\{A_i\} \subseteq \mathcal{G}$. Then let $A'_1 = A_1$ and

$$\begin{aligned} A'_{n+1} &\equiv A_{n+1} \setminus (\cup_{i=1}^n A_i) = A_{n+1} \cap (\cap_{i=1}^n A_i^C) \\ &= \cap_{i=1}^n (A_{n+1} \cap A_i^C) \in \mathcal{G} \end{aligned}$$

because finite intersections of sets of \mathcal{G} are in \mathcal{G} . Since the A'_i are disjoint, it follows $\cup_{i=1}^\infty A_i = \cup_{i=1}^\infty A'_i \in \mathcal{G}$. Therefore, $\mathcal{G} \supseteq \sigma(\mathcal{K})$. ■

Example 32.11.4 Suppose you have (U, \mathcal{F}) and (V, \mathcal{S}) , two measurable spaces. Let $\mathcal{K} \subseteq U \times V$ consist of all sets of the form $A \times B$ where $A \in \mathcal{F}$ and $B \in \mathcal{S}$. This is easily seen to be a π system. When this is done, $\sigma(\mathcal{K})$ is denoted as $\mathcal{F} \times \mathcal{S}$.

Definition 32.11.5 When \mathcal{K} is the open sets of \mathbb{R}^p , the Borel sets, denoted as $\mathcal{B}(\mathbb{R}^p)$, are defined as $\mathcal{B}(\mathbb{R}^p) \equiv \sigma(\mathcal{K})$.

Don't try to describe a typical Borel set. Just use the definition that these are those sets in the smallest σ algebra that contains the open sets. However, if you wanted to give this a try, see Hewitt and Stromberg [20] who do something like this in showing the existence of Lebesgue measurable sets which are not Borel measurable.

For example, here is a useful result about the product of Borel sets.

Lemma 32.11.6 If A_k is a Borel set in \mathbb{R} , then $\prod_{k=1}^p A_k$ is a Borel set in \mathbb{R}^p .

Proof: Let $\pi_k : \mathbb{R}^p \rightarrow \mathbb{R}$ be defined by $\pi_k(x) \equiv x_k$, the k^{th} entry of x . Then

$$\pi_k^{-1}(U) = \mathbb{R} \times \cdots \times \mathbb{R} \times U \times \mathbb{R} \times \cdots \times \mathbb{R}$$

Let \mathcal{G} be those Borel sets B such that $\pi_k^{-1}(B)$ is Borel in \mathbb{R}^p . Then from the above, this is true if B is open. However, it follows from the definition of inverse image that \mathcal{G} is a σ algebra. Therefore, by definition $\mathcal{G} = \mathcal{B}(\mathbb{R}^p)$. Now note that

$$\prod_{k=1}^p A_k = \cap_{k=1}^p \pi_k^{-1}(A_k) \in \mathcal{B}(\mathbb{R}^p). \quad \blacksquare$$

32.12 Product Measures

First of all is a definition.

Definition 32.12.1 Let (X, \mathcal{F}, μ) be a measure space. Then it is called σ finite if there exists an increasing sequence of sets $R_n \in \mathcal{F}$ such that $\mu(R_n) < \infty$ for all n and also $X = \bigcup_{n=1}^{\infty} R_n$.

Now I will show how to define a measure on $\prod_{i=1}^p X_i$ given that $(X_i, \mathcal{F}_i, \mu_i)$ is a σ finite measure space. The main example I have in mind is the case where each $X_i = \mathbb{R}$ and a measure $\mu = m$ to be described a little later, yielding p dimensional Lebesgue measure. However, there is no good reason not to do this in general. It is no harder, so this is what I am doing here.

Let \mathcal{K} denote all subsets of $\mathbf{X} \equiv \prod_{i=1}^p X_i$ which are the form $\prod_{i=1}^p E_i$ where $E_i \in \mathcal{F}_i$. These are called measurable rectangles. Let $\{R_i^n\}_{n=1}^{\infty}$ be the sequence of sets in \mathcal{F}_i whose union is all of X_i , $R_i^n \subseteq R_i^{n+1}$, and $\mu_i(R_i^n) < \infty$. Thus if $\mathbf{R}^n \equiv \prod_{i=1}^p R_i^n$, and $\mathbf{E} \equiv \prod_{i=1}^p E_i$, then

$$\mathbf{R}^n \cap \mathbf{E} = \prod_{i=1}^p R_i^n \cap E_i$$

Let $\mathbf{I} \equiv (i_1, \dots, i_p)$ where (i_1, \dots, i_p) is a permutation of $\{1, \dots, p\}$. Also, to save on space, let \mathbf{F} be a subset of $\prod_{i=1}^p X_i \equiv \mathbf{X}$ and denote the iterated integral

$$\int_{X_{i_1}} \cdots \int_{X_{i_p}} \mathcal{X}_{\mathbf{F}}(x_1, \dots, x_p) d\mu_{i_1} \cdots d\mu_{i_p}$$

as $\int_{\mathbf{I}} \mathcal{X}_{\mathbf{F}}(x_1, \dots, x_p) d\mu_{\mathbf{I}}$. Let \mathcal{G} denote those subsets \mathbf{F} of \mathbf{X} such that all iterated integrals for $\mathcal{X}_{\mathbf{F} \cap \mathbf{R}^n}(x_1, \dots, x_p)$ make sense and are independent of the permutation. Thus, for short,

$$\mathcal{G} \equiv \left\{ \mathbf{F} \subseteq \mathbf{X} : \text{for all } n, \int_{\mathbf{I}} \mathcal{X}_{\mathbf{F} \cap \mathbf{R}^n}(x_1, \dots, x_p) d\mu_{\mathbf{I}} \text{ is independent of } \mathbf{I} \right\}$$

The iterated integral means exactly what the symbols indicate. First you integrate

$$\mathcal{X}_{\mathbf{F}}(x_1, \dots, x_p)$$

with respect to $d\mu_{i_1}$ and then you have a function of the other variables other than x_{i_1} . You then integrate what is left with respect to x_{i_2} and so forth. This is just like what was with iterated integrals in calculus. In order for this to make sense, every function encountered must be measurable with respect to the appropriate σ algebra. Now obviously $\mathcal{K} \subseteq \mathcal{G}$. In fact, if $\mathbf{F} \in \mathcal{K}$, then $\int_{\mathbf{I}} \mathcal{X}_{\mathbf{F} \cap \mathbf{R}^n}(x_1, \dots, x_p) d\mu_{\mathbf{I}} = \prod_{i=1}^p \mu_i(F_i \cap R_i^n)$ for any choice of n .

Proposition 32.12.2 Let \mathcal{K} and \mathcal{G} be as just defined, then $\mathcal{G} \supseteq \sigma(\mathcal{K})$. We define $\sigma(\mathcal{K})$ as \mathcal{F}^p , better denoted as $\mathcal{F}_1 \times \cdots \times \mathcal{F}_p$. Then if

$$\bar{\mu}(\mathbf{F}) \equiv \lim_{n \rightarrow \infty} \int_{\mathbf{I}} \mathcal{X}_{\mathbf{F} \cap \mathbf{R}^n}(x_1, \dots, x_p) d\mu_{\mathbf{I}},$$

then $\bar{\mu}$ is a measure which does not depend on \mathbf{I} the particular permutation chosen for the order of integration. $\bar{\mu}$ often denoted as $\mu_1 \times \cdots \times \mu_p$ is called product measure.

$f : \mathbf{X} \rightarrow [0, \infty)$ is measurable with respect to \mathcal{F}^p then for any permutation (i_1, \dots, i_p) of $\{1, \dots, p\}$ it follows

$$\int f d\vec{\mu} = \int \cdots \int f(x_1, \dots, x_p) d\mu_{i_1} \cdots d\mu_{i_p} \quad (32.14)$$

Proof: I will show that \mathcal{G} is closed with respect to complements and countable disjoint unions. Then the result will follow. Now suppose $\{\mathbf{F}^k\}_{k=1}^\infty$ are disjoint, each in \mathcal{G} . Then if $\mathbf{F} \equiv \bigcup_{k=1}^\infty \mathbf{F}^k$, $\mathbf{F} \cap \mathbf{R}^n = \bigcup_{k=1}^\infty \mathbf{F}^k \cap \mathbf{R}^n$ and since these sets are disjoint,

$$\mathcal{X}_{\mathbf{F} \cap \mathbf{R}^n} = \sum_{k=1}^\infty \mathcal{X}_{\mathbf{F}^k \cap \mathbf{R}^n}$$

Therefore, applying the monotone convergence theorem repeatedly for the iterated integrals and using the fact that measurability is not lost on taking limits, then for (i_1, \dots, i_p) , (j_1, \dots, j_p) two permutations,

$$\begin{aligned} & \int \cdots \int \mathcal{X}_{\mathbf{F} \cap \mathbf{R}^n}(x_1, \dots, x_p) d\mu_{i_1} \cdots d\mu_{i_p} \\ &= \int \cdots \int \lim_{N \rightarrow \infty} \sum_{k=1}^N \mathcal{X}_{\mathbf{F}^k \cap \mathbf{R}^n}(x_1, \dots, x_p) d\mu_{i_1} \cdots d\mu_{i_p} \\ &= \lim_{N \rightarrow \infty} \int \cdots \int \sum_{k=1}^N \mathcal{X}_{\mathbf{F}^k \cap \mathbf{R}^n}(x_1, \dots, x_p) d\mu_{i_1} \cdots d\mu_{i_p} \\ &= \lim_{N \rightarrow \infty} \sum_{k=1}^N \int \cdots \int \mathcal{X}_{\mathbf{F}^k \cap \mathbf{R}^n}(x_1, \dots, x_p) d\mu_{i_1} \cdots d\mu_{i_p} \\ &= \lim_{N \rightarrow \infty} \sum_{k=1}^N \int \cdots \int \mathcal{X}_{\mathbf{F}^k \cap \mathbf{R}^n}(x_1, \dots, x_p) d\mu_{j_1} \cdots d\mu_{j_p} \\ &= \lim_{N \rightarrow \infty} \int \cdots \int \sum_{k=1}^N \mathcal{X}_{\mathbf{F}^k \cap \mathbf{R}^n}(x_1, \dots, x_p) d\mu_{j_1} \cdots d\mu_{j_p} \\ &= \int \cdots \int \lim_{N \rightarrow \infty} \sum_{k=1}^N \mathcal{X}_{\mathbf{F}^k \cap \mathbf{R}^n}(x_1, \dots, x_p) d\mu_{j_1} \cdots d\mu_{j_p} \\ &= \int \cdots \int \mathcal{X}_{\mathbf{F} \cap \mathbf{R}^n}(x_1, \dots, x_p) d\mu_{j_1} \cdots d\mu_{j_p} \end{aligned}$$

Thus \mathcal{G} is closed with respect to countable disjoint unions. So suppose $\mathbf{F} \in \mathcal{G}$. Then $\mathcal{X}_{\mathbf{F}^c \cap \mathbf{R}^n} = \mathcal{X}_{\mathbf{R}^n} - \mathcal{X}_{\mathbf{F} \cap \mathbf{R}^n}$. Everything works for both terms on the right and in addition, $\int_{\mathbf{I}} \mathcal{X}_{\mathbf{R}^n} d\mu_{\mathbf{I}}$ is finite and independent of \mathbf{I} . Therefore, everything works as it should for the function on the left using similar arguments to the above. You simply verify that all makes sense for each integral at a time and apply monotone convergence theorem as needed. Therefore, \mathcal{G} is indeed closed with respect to complements. It follows that $\mathcal{G} \supseteq \sigma(\mathcal{K})$ by Dynkin's lemma, Lemma 32.11.3. Now define for $\mathbf{F} \in \sigma(\mathcal{K})$,

$$\vec{\mu}(\mathbf{F}) \equiv \lim_{n \rightarrow \infty} \int_{\mathbf{I}} \mathcal{X}_{\mathbf{F} \cap \mathbf{R}^n}(x_1, \dots, x_p) d\mu_{\mathbf{I}}$$

By definition of \mathcal{G} this definition of $\bar{\mu}$ does not depend on I . If you have $\{F^k\}_{k=1}^\infty$ is a sequence of disjoint sets in \mathcal{G} , then if F is their union,

$$\bar{\mu}(F) \equiv \lim_{n \rightarrow \infty} \int_I \sum_{k=1}^n \mathcal{X}_{F^k \cap R^n}(x_1, \dots, x_p) d\mu_I$$

and one can apply the monotone convergence theorem one integral at a time and obtain that this is

$$\lim_{n \rightarrow \infty} \sum_{k=1}^n \int_I \mathcal{X}_{F^k \cap R^n}(x_1, \dots, x_p) d\mu_I$$

Now applying the monotone convergence theorem again, this time for the Lebesgue integral given by a sum with counting measure, the above is

$$\sum_{k=1}^{\infty} \lim_{n \rightarrow \infty} \int_I \mathcal{X}_{F^k \cap R^n}(x_1, \dots, x_p) d\mu_I \equiv \sum_{k=1}^{\infty} \bar{\mu}(F^k)$$

which shows that $\bar{\mu}$ is indeed a measure. Also from the construction, it follows that this measure does not depend on the particular permutation of the iterated integrals used to compute it.

The claim about the integral 32.14 follows right away from the monotone convergence theorem applied in the right side one iterated integral at a time and approximation with simple functions as in Theorem 32.2.8. The result holds for each of an increasing sequence simple functions from linearity of integrals and the definition of $\bar{\mu}$. Then you apply the monotone convergence theorem to obtain the claim of the theorem. ■

32.13 Exercises

1. Show carefully that if \mathfrak{S} is a set whose elements are σ algebras which are subsets of $\mathcal{P}(\Omega)$, then $\cap \mathfrak{S}$ is also a σ algebra. Now let $\mathcal{G} \subseteq \mathcal{P}(\Omega)$ satisfy property P if \mathcal{G} is closed with respect to complements and countable disjoint unions as in Dynkin's lemma, and contains \emptyset and Ω . If $\mathfrak{H} \subseteq \mathcal{G}$ is any set whose elements are subsets of $\mathcal{P}(\Omega)$ which satisfies property P , then $\cap \mathfrak{H}$ also satisfies property P . Thus there is a smallest subset of \mathcal{G} satisfying P .
2. Show $\mathcal{B}(\mathbb{R}^p) = \sigma(\mathcal{P})$ where \mathcal{P} consists of the half open rectangles which are of the form $\prod_{i=1}^p [a_i, b_i)$. Recall $\mathcal{B}(\mathbb{R}^p)$ is the smallest σ algebra containing the open sets.
3. Show that $f : (\Omega, \mathcal{F}) \rightarrow \mathbb{R}$ is measurable if and only if $f^{-1}(\text{open}) \in \mathcal{F}$. Show that if E is any set in $\mathcal{B}(\mathbb{R})$, then $f^{-1}(E) \in \mathcal{F}$. Thus, inverse images of Borel sets are measurable. Next consider $f : (\Omega, \mathcal{F}) \rightarrow \mathbb{R}$ being measurable and $g : \mathbb{R} \rightarrow \mathbb{R}$ is Borel measurable, meaning that $g^{-1}(\text{open}) \in \mathcal{B}(\mathbb{R})$. Explain why $g \circ f$ is measurable. **Hint:** You know that $(g \circ f)^{-1}(U) = f^{-1}(g^{-1}(U))$. 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. However, these things are not for this book.

4. Let $X_i \equiv \mathbb{R}^{n_i}$ and let $X = \prod_{i=1}^n X_i$ and let the distance between two points in X be given by

$$\|x - y\| \equiv \max \{\|x_i - y_i\|, i = 1, 2, \dots, n\}$$

Show that any set of the form

$$\prod_{i=1}^n E_i, E_i \in \mathcal{B}(X_i)$$

is a Borel set. That is, the product of Borel sets is Borel. **Hint:** You might consider the continuous functions $\pi_i : \prod_{j=1}^n X_j \rightarrow X_i$ which are the projection maps. Thus $\pi_i(x) \equiv x_i$. Then $\pi_i^{-1}(E_i)$ would have to be Borel measurable whenever $E_i \in \mathcal{B}(X_i)$. Explain why. You know π_i is continuous. Why would $\pi_i^{-1}(\text{Borel})$ be a Borel set? Then you might argue that $\prod_{i=1}^n E_i = \cap_{i=1}^n \pi_i^{-1}(E_i)$. Set the text for a special case that $X_i = \mathbb{R}$ in the next chapter.

5. You have two finite measures defined on $\mathcal{B}(X)$ 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. Let $\mu(E) = 1$ if $0 \in E$ and $\mu(E) = 0$ if $0 \notin E$. Show this is a measure on $\mathcal{P}(\mathbb{R})$.
7. Give an example of a measure μ and a measure space and a decreasing sequence of measurable sets $\{E_i\}$ such that $\lim_{n \rightarrow \infty} \mu(E_n) \neq \mu(\cap_{i=1}^{\infty} E_i)$.
8. If you have a finite measure μ on $\mathcal{B}(\mathbb{R}^p)$, and if $E \in \mathcal{B}(\mathbb{R}^p)$, show that there exist sets F, G such that G is a countable intersection of open sets, called a G_δ set and F a countable union of closed sets, called an F_σ set, such that $F \subseteq E \subseteq G$ and $\mu(G \setminus F) = 0$. **Hint:** Show first for G open. Then you might try to use Dynkin's lemma to extend to Borel sets. Recall that $\mathcal{B}(\mathbb{R}^p) = \sigma(\mathcal{K})$ where \mathcal{K} consists of the open sets. In the first part, you might want to use Proposition 15.6.4 to produce an increasing sequence closed sets whose union is the open set G .
9. You have a measure space (Ω, \mathcal{F}, P) where P is a probability measure on \mathcal{F} . Then you also have a measurable function $X : \Omega \rightarrow \mathbb{R}^p$, meaning that $X^{-1}(U) \in \mathcal{F}$ whenever U is open. Now define a measure on $\mathcal{B}(\mathbb{R}^p)$ 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 $\mathcal{B}(\mathbb{R}^n)$. This is called the distribution measure.
10. Let $K \subseteq V$ where K is closed and V is open. Consider the following function.

$$f(x) = \frac{\text{dist}(x, V^C)}{\text{dist}(x, K) + \text{dist}(x, V^C)}$$

Explain why this function is continuous, equals 0 off V and equals 1 on K . The needed function is in Proposition 15.6.4.

11. Let (Ω, \mathcal{F}) be a measurable space and let $f : \Omega \rightarrow \mathbb{R}^p$ be a measurable function meaning that $f^{-1}(U) \in \mathcal{F}$ whenever U is open. Then $\sigma(f)$ denotes the smallest σ algebra such that f is measurable with respect to this σ algebra. Show that $\sigma(f) = \{f^{-1}(E) : E \in \mathcal{B}(\mathbb{R}^p)\}$.

12. There is a monumentally important theorem called the Borel Cantelli lemma. It says the following. If you have a measure space $(\Omega, \mathcal{F}, \mu)$ and if $\{E_i\} \subseteq \mathcal{F}$ is such that $\sum_{i=1}^{\infty} \mu(E_i) < \infty$, 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 . **Hint:** You might look at the set of all ω which are in infinitely many of the E_i . First explain why this set is of the form $\bigcap_{n=1}^{\infty} \bigcup_{k \geq n} E_k$.
13. Let $(\Omega, \mathcal{F}, \mu)$ 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 $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \mu(\{\omega : |f_n(\omega) - f(\omega)| > \varepsilon\}) = 0$$

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

$$\lim_{n_k \rightarrow \infty} f_{n_k}(\omega) = f(\omega).$$

Also show that if μ is finite and $\lim_{n \rightarrow \infty} f_n(\omega) = f(\omega)$, then f_n converges in measure to f .

14. Let $\{r_n\}_{n=1}^{\infty}$ be an enumeration of the rational numbers in $[0, 1]$ meaning that every rational number is included in $\{r_n\}_{n=1}^{\infty}$ for some n and let $f_n(x) = 0$ except for when $x \in \{r_1, \dots, r_n\}$ when it is 1. Explain why f_n is Riemann integrable and has Riemann integral 0. However, $\lim_{n \rightarrow \infty} f_n(x) \equiv f(x)$ is 1 on rationals and 0 elsewhere so this isn't even Riemann integrable. It will be shown later that the two integrals give the same answer whenever the function is Riemann integrable. Thus the Lebesgue integral of f_n will be 0. So what is the Lebesgue integral of the function which is 1 on the rationals and 0 on the irrationals? Explain why this is so.
15. Prove Chebyshev's inequality $m_p(\{x : |f(x)| > \lambda\}) \leq \frac{1}{\lambda} \|f\|_{L^1} \equiv \frac{1}{\lambda} \int |f| dm_p$.

Chapter 33

The Lebesgue Measure and Integral in \mathbb{R}^p

33.1 An Outer Measure on $\mathcal{P}(\mathbb{R})$

It is needed to find a measure which delivers length. Recall $\mathcal{P}(S)$ denotes the set of all subsets of S . To begin with, it is shown there is an outer measure which gives length.

Theorem 33.1.1 *There exists a function $m : \mathcal{P}(\mathbb{R}) \rightarrow [0, \infty]$ which satisfies the following properties.*

1. If $A \subseteq B$, then $0 \leq m(A) \leq m(B)$, $m(\emptyset) = 0$.
2. $m(\cup_{k=1}^{\infty} A_k) \leq \sum_{i=1}^{\infty} m(A_i)$
3. $m([a, b]) = b - a = m((a, b))$.

Proof: First it is necessary to define the function m . This is contained in the following definition.

Definition 33.1.2 For $A \subseteq \mathbb{R}$,

$$m(A) = \inf \left\{ \sum_{i=1}^{\infty} (b_i - a_i) : A \subseteq \cup_{i=1}^{\infty} (a_i, b_i) \right\}$$

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 $m(\emptyset) = 0$? Then $\emptyset \subseteq (a - \delta, a + \delta)$ and so $m(\emptyset) \leq 2\delta$ for every $\delta > 0$. Letting $\delta \rightarrow 0$, it follows that $m(\emptyset) = 0$.

Consider 2.). If any $m(A_i) = \infty$, there is nothing to prove. The assertion simply is $\infty \leq \infty$. Assume then that $m(A_i) < \infty$ for all i . Then for each $m \in \mathbb{N}$ there exists a countable

set of open intervals, $\{(a_i^m, b_i^m)\}_{i=1}^\infty$ whose union contains A_m such that

$$m(A_m) + \frac{\varepsilon}{2^m} > \sum_{i=1}^\infty (b_i^m - a_i^m).$$

Then using Theorem 6.6.4 on Page 175,

$$\begin{aligned} m(\cup_{m=1}^\infty A_m) &\leq \sum_{i,m} (b_i^m - a_i^m) = \sum_{m=1}^\infty \sum_{i=1}^\infty (b_i^m - a_i^m) \\ &\leq \sum_{m=1}^\infty m(A_m) + \frac{\varepsilon}{2^m} = \sum_{m=1}^\infty m(A_m) + \varepsilon, \end{aligned}$$

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 $m([a, b]) + \varepsilon \geq \sum_{i=1}^\infty (b_i - a_i)$. Since $[a, b]$ is compact, finitely many of these 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 $a \in (a_1, b_1), b_1 \in (a_2, b_2), \dots, b \in (a_n, b_n)$. Therefore, $m([a, b]) \leq \sum_{i=1}^n (b_i - a_i)$. It follows

$$\sum_{i=1}^n (b_i - a_i) \geq m([a, b]) \geq \sum_{i=1}^n (b_i - a_i) - \varepsilon \geq (b - a) - \varepsilon$$

Therefore, since $(a - \frac{\varepsilon}{2}, b + \frac{\varepsilon}{2}) \supseteq [a, b]$,

$$(b - a) + \varepsilon \geq m([a, b]) \geq (b - a) - \varepsilon$$

Since ε is arbitrary, $(b - a) = m([a, b])$. Consider $[a + \delta, b - \delta]$. From what was just shown, $m([a + \delta, b - \delta]) = (b - a) - 2\delta \leq m([a, b])$ and so, since this holds for every δ , $(b - a) \leq m((a, b)) \leq m([a, b]) = (b - a)$. This shows 3.) ■

33.2 One Dimensional Lebesgue Measure

Theorem 33.2.1 *Let \mathcal{F} denote the σ algebra of Theorem 32.4.4, associated with the outer measure m in Theorem 33.1.1, on which m is a measure. Then every open interval is in \mathcal{F} . All open sets are in \mathcal{F} and all half open and closed intervals are in \mathcal{F} .*

Proof: The first task is to show $(a, b) \in \mathcal{F}$. I need to show that for every $S \subseteq \mathbb{R}$,

$$m(S) \geq m(S \cap (a, b)) + m(S \cap (a, b)^c) \quad (33.1)$$

Suppose first S is an open interval, (c, d) . If (c, d) has empty intersection with (a, b) or is contained in (a, b) there is nothing to prove. The above expression reduces to nothing more than $m(S) = m(S)$. Suppose next that $(c, d) \supseteq (a, b)$. In this case, the right side of the above reduces to

$$\begin{aligned} &m((a, b)) + m((c, a] \cup [b, d)) \leq b - a + a - c + d - b \\ &= d - c = m((c, d)) \end{aligned}$$

The only other cases are $c \leq a < d \leq b$ or $a \leq c < d \leq b$. Consider the first of these cases. Then the right side of 33.1 for $S = (c, d)$ is

$$m((a, d)) + m((c, a]) = d - a + a - c = m((c, d))$$

The last case is entirely similar. Thus 33.1 holds whenever S is an open interval. Now it is clear 33.1 also holds if $m(S) = \infty$. Suppose then that $m(S) < \infty$ and let

$$S \subseteq \bigcup_{k=1}^{\infty} (a_k, b_k)$$

such that

$$m(S) + \varepsilon > \sum_{k=1}^{\infty} (b_k - a_k) = \sum_{k=1}^{\infty} m((a_k, b_k)).$$

Then since m is an outer measure, and using what was just shown,

$$\begin{aligned} & m(S \cap (a, b)) + m\left(S \cap (a, b)^C\right) \\ & \leq m\left(\bigcup_{k=1}^{\infty} (a_k, b_k) \cap (a, b)\right) + m\left(\bigcup_{k=1}^{\infty} (a_k, b_k) \cap (a, b)^C\right) \\ & \leq \sum_{k=1}^{\infty} m\left((a_k, b_k) \cap (a, b)\right) + m\left((a_k, b_k) \cap (a, b)^C\right) \\ & \leq \sum_{k=1}^{\infty} m((a_k, b_k)) \leq m(S) + \varepsilon. \end{aligned}$$

Since ε is arbitrary, this shows 33.1 holds for any S and so any open interval is in \mathcal{F} . By Theorem 32.1.5, every open set is a countable union of open intervals. Therefore, all open sets are in \mathcal{F} . As to half open intervals, $(a, b] = \bigcap_{n=1}^{\infty} \left(a, b + \frac{1}{n}\right) = \left(\bigcup_{n=1}^{\infty} \left(a, b + \frac{1}{n}\right)\right)^C \in \mathcal{F}$. A similar argument shows that closed intervals are in \mathcal{F} also. ■

33.3 The Lebesgue Integral and Riemann Integral

How does the Lebesgue integral taken with respect to Lebesgue measure compare with the one dimensional Riemann integral of a nonnegative continuous bounded function? First of all, to save space, I will write $\int_a^b f dm$ for the Lebesgue integral $\int \mathcal{X}_{[a,b]} f dm$. The following proposition shows that when a function is Riemann integrable, it is also Lebesgue integrable and the two integrals give the same answer.

Proposition 33.3.1 *Let $f \geq 0$ and let it be in $R([a, b])$. Then f is Lebesgue integrable and*

$$\int_a^b f(x) dx = \int_a^b f dm$$

Proof: By the Riemann criterion, there exist upper sums $U(f, P_n)$ and lower sums $L(f, P_n)$ such that $U(f, P_n) - L(f, P_n) < 2^{-n}$. Let $a_n(x)$ be a step function corresponding to $U(f, P_n)$ such that $\int_a^b a_n dm = \int_a^b a_n(x) dx$ and let $b_n(x)$ be a step function corresponding to $L(f, P_n)$, $\int_a^b b_n dm = \int_a^b b_n(x) dx$. Thus $b_n(x) \leq f(x) \leq a_n(x)$. We can also arrange to have the partitions be increasing so that $b_n(x) \leq b_{n+1}(x) \cdots$, $a_n(x) \geq a_{n+1}(x) \cdots$. These step functions are constant on intervals or half open intervals. Now every interval is a Borel

set (Why?) and so these functions are Borel measurable. Let $g(x) \equiv \lim_{n \rightarrow \infty} b_n(x)$, $h(x) \equiv \lim_{n \rightarrow \infty} a_n(x)$. Then $g(x) \leq f(x) \leq h(x)$ and $\int_a^b (h - g) dm = 0$. Therefore, off a set of measure zero $h(x) = g(x)$. By completeness of Lebesgue measure, it follows that f must be Lebesgue measurable because it is not equal to the Borel function g only on a subset of the set of measure zero where $h(x) \neq g(x)$. Also, by the monotone convergence theorem,

$$\begin{aligned} \int_a^b f dm &= \int_a^b g dm = \lim_{n \rightarrow \infty} \int_a^b b_n dm = \lim_{n \rightarrow \infty} \int_a^b b_n(x) dx \\ &= \lim_{n \rightarrow \infty} L(f, P_n) = \int_a^b f(x) dx \blacksquare \end{aligned}$$

What if f is bounded, continuous but maybe not nonnegative? Then you can write $f = f_+ - f_-$ where, as before, $f_+ \equiv \frac{|f|+f}{2}$, $f_- \equiv \frac{|f|-f}{2}$. These $x \rightarrow x_+$, $x \rightarrow x_-$ are continuous and so f_+ , f_- are measurable. You know that

$$\int_a^b f dx = \int_a^b f_+ dx - \int_a^b f_- dx = \int_a^b f_+ dm - \int_a^b f_- dm \equiv \int_a^b f dm$$

Theorem 33.3.2 *If $f \in R([a, b])$, then the Riemann and Lebesgue integral are the same. Thus you can apply the fundamental theorem of calculus to compute the integral.*

33.4 p Dimensional Lebesgue Measure and Integrals

33.4.1 Iterated Integrals

Let m denote one dimensional Lebesgue measure. Also let the σ algebra of measurable sets be denoted by \mathcal{F} . Recall this σ algebra contained the open sets. Also from the construction given above,

$$m([a, b]) = m((a, b)) = b - a$$

Definition 33.4.1 *Let f be a function of p variables and consider the s symbol*

$$\int \cdots \int f(x_1, \dots, x_p) dx_{i_1} \cdots dx_{i_p}. \quad (33.2)$$

where (i_1, \dots, i_p) is a permutation of the integers $\{1, 2, \dots, p\}$. The symbol means to first do the Lebesgue integral

$$\int f(x_1, \dots, x_p) dx_{i_1}$$

yielding a function of the other $p - 1$ variables given above. Then you do

$$\int \left(\int f(x_1, \dots, x_p) dx_{i_1} \right) dx_{i_2}$$

and continue this way. The iterated integral is said to make sense if the process just described makes sense at each step. Thus, to make sense, it is required

$$x_{i_1} \rightarrow f(x_1, \dots, x_p)$$

can be integrated. Either the function has values in $[0, \infty]$ and is measurable or it is a function in L^1 . Then it is required

$$x_{i_2} \rightarrow \int f(x_1, \dots, x_p) dx_{i_1}$$

can be integrated and so forth. The symbol in 33.2 is called an iterated integral.

With the above explanation of iterated integrals, it is now time to define p dimensional Lebesgue measure.

33.4.2 p Dimensional Lebesgue Measure and Integrals

Consider $(\mathbb{R}, \mathcal{F}, m)$ the measure space corresponding to one dimensional Lebesgue measure. Then from Proposition 32.12.2, we obtain the existence of p dimensional Lebesgue measure.

Proposition 33.4.2 *There exists a measure m_p defined on \mathcal{F}^p such that if $f: \mathbb{R}^p \rightarrow [0, \infty]$ is measurable with respect to \mathcal{F}^p then for any permutation (i_1, \dots, i_p) of $\{1, \dots, p\}$ it follows*

$$\int_{\mathbb{R}^p} f dm_p = \int \cdots \int f(x_1, \dots, x_p) dx_{i_1} \cdots dx_{i_p} \quad (33.3)$$

In particular, this implies that if A_i is a Borel set for each $i = 1, \dots, p$ then

$$m_p \left(\prod_{i=1}^p A_i \right) = \prod_{i=1}^p m(A_i).$$

and all such $\prod_{i=1}^p A_i$ is in \mathcal{F}^p .

This will suffice for this book. Actually, you use the completion of this measure space and this completion is Lebesgue measure. Writing such a measurable function as a difference between positive and negative parts, gives the following corollary.

Corollary 33.4.3 *In the context of the above Proposition 33.4.2, if $f \in L^1(\mathbb{R}^p)$, then for any permutation (i_1, \dots, i_p) of $\{1, \dots, p\}$ it follows*

$$\int_{\mathbb{R}^p} f dm_p = \int \cdots \int f(x_1, \dots, x_p) dx_{i_1} \cdots dx_{i_p} \quad (33.4)$$

Note that this implies that if $\int \cdots \int |f(x_1, \dots, x_p)| dx_{i_1} \cdots dx_{i_p} < \infty$, the integration taken in any order, then 33.4 holds for all permutations.

The next big theorem about the integral is the change of variables formula. Recall Lemma 32.1.6.

Lemma 33.4.4 *Every open set in \mathbb{R}^p is the countable disjoint union of half open boxes of the form*

$$\prod_{i=1}^p (a_i, a_i + 2^{-k}]$$

where $a_i = l2^{-k}$ for some integers, l, k where $k \geq m$. If \mathcal{B}_m denotes this collection of half open boxes, then every box of \mathcal{B}_{m+1} is contained in a box of \mathcal{B}_m or equals a box of \mathcal{B}_m .

33.5 Lebesgue Measure and Linear Maps

Lemma 33.5.1 *Let $A : \mathbb{R}^p \rightarrow \mathbb{R}^p$ be linear and invertible. Then A maps open sets to open sets.*

Proof: This follows from the observation that if B is any linear transformation, then B is continuous. Indeed, it is realized by matrix multiplication and so it is clear that if $x_n \rightarrow x$, then $Bx_n \rightarrow Bx$. Now it follows that A^{-1} is continuous. Let U be open. Let $y \in A(U)$. Then is y an interior point of $A(U)$? if not, there exists $y_n \rightarrow y$ where $y_n \notin A(U)$. But then $A^{-1}y_n \rightarrow A^{-1}y \in U$. Since U is open, $A^{-1}y_n \in U$ for all n large enough and so $y_n \in A(U)$ after all. Thus y is an interior point of $A(U)$ showing that $A(U)$ is open. ■

Corollary 33.5.2 *Let $A : \mathbb{R}^p \rightarrow \mathbb{R}^p$ be linear and invertible. Then A maps Borel sets to Borel sets.*

Proof: Let the pi system be \mathcal{K} the open sets. Then let \mathcal{G} be those Borel sets E such that $A(E)$ is Borel. Then it is clear that \mathcal{G} contains \mathcal{K} and is closed with respect to complements and countable disjoint unions. By Dynkin's lemma, $\mathcal{G} = \mathcal{B}(\mathbb{R}^p) = \sigma(\mathcal{K})$. This last equality holds by definition of the Borel sets $\mathcal{B}(\mathbb{R}^p)$. ■

From Linear algebra, Chapter 18 the chapter on row operations and elementary matrices, if A is such an invertible linear transformation, it is the composition of finitely many invertible linear transformations which are of the following form.

$$\begin{aligned} & \begin{pmatrix} x_1 & \cdots & x_r & \cdots & x_s & \cdots & x_p \end{pmatrix}^T \\ & \rightarrow \begin{pmatrix} x_1 & \cdots & x_r & \cdots & x_s & \cdots & x_p \end{pmatrix}^T \\ & \begin{pmatrix} x_1 & \cdots & x_r & \cdots & x_p \end{pmatrix}^T \rightarrow \begin{pmatrix} x_1 & \cdots & cx_r & \cdots & x_p \end{pmatrix}^T, c \neq 0 \\ & \begin{pmatrix} x_1 & \cdots & x_r & \cdots & x_s & \cdots & x_p \end{pmatrix}^T \\ & \rightarrow \begin{pmatrix} x_1 & \cdots & x_r & \cdots & x_s + x_r & \cdots & x_p \end{pmatrix}^T \end{aligned}$$

where these are the actions obtained by multiplication by elementary matrices. Denote these special linear transformations by $E(r \leftrightarrow s)$, $E(cr)$, $E(s \rightarrow s+r)$.

Let $R = \prod_{i=1}^p (a_i, b_i)$. Then it is easily seen that

$$\begin{aligned} m_p(E(r \leftrightarrow s)(R)) &= m_p(R) = |\det(E(r \leftrightarrow s))| m_p(R) \\ m_p(E(cr)(R)) &= |c| m_p(R) = |\det(E(cr))| m_p(R) \end{aligned}$$

The other linear transformation which represents a sheer is a little harder. However,

$$\begin{aligned} m_p(E(s \rightarrow s+r)(R)) &= \int_{E(s \rightarrow s+r)(R)} dm_p \\ &= \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \int_{\mathbb{R}} \int_{\mathbb{R}} \mathcal{X}_{E(s \rightarrow s+r)(R)} dx_s dx_r dx_{p_1} \cdots dx_{p_{p-2}} \end{aligned}$$

Now recall Theorem 33.3.2 which says you can integrate using the usual Riemann integral when the function involved is continuous. Thus the above becomes

$$\begin{aligned} & \int_{a_{p_{p-2}}}^{b_{p_{p-2}}} \cdots \int_{a_{p_1}}^{b_{p_1}} \int_{a_r}^{b_r} \int_{a_s+x_r}^{b_s+x_r} dx_s dx_r dx_{p_1} \cdots dx_{p_{p-2}} \\ &= m_p(R) = |\det(E(s \rightarrow s+r))| m_p(R) \end{aligned}$$

Recall that when a row (column) is added to another row (column), the determinant of the resulting matrix is unchanged.

Lemma 33.5.3 *Let L be any of the above elementary linear transformations. Then*

$$m_p(L(F)) = |\det(L)| m_p(F)$$

for any Borel set F . Also $L(F)$ is Borel if F is Borel.

Proof: Let $R_k = \prod_{i=1}^p (-k, k)$. Let \mathcal{G} be those Borel sets F such that

$$m_p(L(F \cap R_k)) = |\det(L)| m_p(F \cap R_k) \quad (33.5)$$

Letting \mathcal{H} be the open boxes, it follows from the above discussion that the pi system \mathcal{H} is in \mathcal{G} . It is also obvious that if $F_i \in \mathcal{G}$ the F_i being disjoint, then

$$\begin{aligned} m_p(L(\cup_{i=1}^{\infty} F_i \cap R_k)) &= \sum_{i=1}^{\infty} m_p(L(F_i \cap R_k)) = |\det(L)| \sum_{i=1}^{\infty} m_p(F_i \cap R_k) \\ &= |\det(L)| m_p(\cup_{i=1}^{\infty} F_i \cap R_k) \end{aligned}$$

Thus \mathcal{G} is closed with respect to countable disjoint unions. If $F \in \mathcal{G}$ then

$$m_p(L(F^C \cap R_k)) + m_p(L(F \cap R_k)) = m_p(L(R_k))$$

$$m_p(L(F^C \cap R_k)) + |\det(L)| m_p(F \cap R_k) = |\det(L)| m_p(R_k)$$

$$\begin{aligned} m_p(L(F^C \cap R_k)) &= |\det(L)| m_p(R_k) - |\det(L)| m_p(F \cap R_k) \\ &= |\det(L)| m_p(F^C \cap R_k) \end{aligned}$$

It follows that \mathcal{G} is closed with respect to complements also. Therefore, $\mathcal{G} = \sigma(\mathcal{H}) = \mathcal{B}(\mathbb{R}^p)$. Now let $k \rightarrow \infty$ in 33.5 to obtain the desired conclusion. ■

Theorem 33.5.4 *Let L be a linear transformation which is invertible. Then for any Borel F , $L(F)$ is Borel and*

$$m_p(L(F)) = |\det(L)| m_p(F)$$

Proof: From linear algebra, there are L_i each elementary such that $L = L_1 \circ L_2 \circ \cdots \circ L_s$. By Corollary 33.5.2, each L_i maps Borel sets to Borel sets. Hence, using Lemma 33.5.3

$$\begin{aligned} m_p(L(F)) &= |\det(L_1)| m_p(L_2 \circ \cdots \circ L_s(F)) \\ &= |\det(L_1)| |\det(L_2)| m_p(L_3 \circ \cdots \circ L_s(F)) \\ &= \cdots = \prod_{i=1}^s |\det(L_i)| m_p(F) = |\det(L)| m_p(F) \end{aligned}$$

the last claim from properties of the determinant. ■

33.6 Change of Variables for Nonlinear Maps

Assume the following:

1. $V = \mathbf{h}(U)$, U, V open and bounded, \mathbf{h} one to one.
2. $\mathbf{h}, \mathbf{h}^{-1}$ are $C^1(\hat{U}), C^1(\hat{V})$ respectively where $\hat{U} \supseteq \bar{U}, \hat{V} \supseteq \bar{V}$.

Let the balls be defined in terms of the norm

$$\|\mathbf{x}\| \equiv \max \{|x_k| : k = 1, \dots, p\}$$

Note that $|\mathbf{x}| \geq \|\mathbf{x}\| \geq \frac{1}{\sqrt{p}} |\mathbf{x}|$ so it doesn't matter which norm you use in the definition of differentiability. $\|\cdot\|$ happens to be a little more convenient here.

Then define

$$\phi(\mathbf{x}, \mathbf{v}) \equiv \frac{\|\mathbf{h}(\mathbf{x} + \mathbf{v}) - (\mathbf{h}(\mathbf{x}) + D\mathbf{h}(\mathbf{x})\mathbf{v})\|}{\|\mathbf{v}\|} \quad (33.6)$$

Then ϕ is continuous on $\bar{U} \times \overline{B(\mathbf{0}, 1)}$ with the convention that $\phi(\mathbf{x}, \mathbf{0}) \equiv 0$. Thus it is uniformly continuous on this compact set and so there exists $\delta > 0$ such that if $\|\mathbf{v}\| < \delta$, then

$$|\phi(\mathbf{x}, \mathbf{v}) - \phi(\mathbf{x}, \mathbf{0})| = |\phi(\mathbf{x}, \mathbf{v})| < \varepsilon, \quad (33.7)$$

this for all $\mathbf{x} \in \bar{U}$.

$$\begin{aligned} \mathbf{h}(\mathbf{x} + \mathbf{v}) - \mathbf{h}(\mathbf{x}) &= D\mathbf{h}(\mathbf{x})\mathbf{v} + \mathbf{o}(\mathbf{v}) \\ &= D\mathbf{h}(\mathbf{x})(\mathbf{v} + D\mathbf{h}^{-1}(\mathbf{h}(\mathbf{x}))\mathbf{o}(\mathbf{v})) \end{aligned}$$

Let $f : V \rightarrow \mathbb{R}$ be a bounded, uniformly continuous function.

Let \mathcal{B}_m be a collection of disjoint half open rectangles as in Lemma 32.1.6 such that each has diameter no more than 2^{-m} and each rectangle of \mathcal{B}_{m+1} is either a subset of a rectangle of \mathcal{B}_m or is equal to a rectangle of \mathcal{B}_m such that $\cup \mathcal{B}_m = U$. Let m be large enough that the diameters of all these half open rectangles are less than δ . Denote the rectangles of \mathcal{B}_m as $\{R_i^m\}_{i=1}^\infty$ and let the center of these be denoted by \mathbf{x}_i^m . Also let m be large enough that

$$|f(\mathbf{h}(\mathbf{x}_i^m))| \det(D\mathbf{h}(\mathbf{x}_i^m)) - f(\mathbf{h}(\mathbf{x})) |\det(D\mathbf{h}(\mathbf{x}))| < \varepsilon \text{ for all } \mathbf{x} \in R_i^m$$

A basic version of the theorems to be presented is the following.

Lemma 33.6.1 *Let U and V be bounded open sets in \mathbb{R}^p and let $\mathbf{h}, \mathbf{h}^{-1}$ be C^1 defined respectively on $\hat{U} \supseteq \bar{U}$ and $\hat{V} \supseteq \bar{V}$ such that $\mathbf{h}(U) = V$ and let f be a bounded uniformly continuous function defined on U . Then*

$$\int_V f(\mathbf{y}) dm_p = \int_U f(\mathbf{h}(\mathbf{x})) |\det(D\mathbf{h}(\mathbf{x}))| dm_p$$

Proof: Let $\mathbf{x} \in U$. By the assumption that \mathbf{h} and \mathbf{h}^{-1} are C^1 ,

$$\begin{aligned} \mathbf{h}(\mathbf{x} + \mathbf{v}) - \mathbf{h}(\mathbf{x}) &= D\mathbf{h}(\mathbf{x})\mathbf{v} + \mathbf{o}(\mathbf{v}) \\ &= D\mathbf{h}(\mathbf{x})(\mathbf{v} + D\mathbf{h}^{-1}(\mathbf{h}(\mathbf{x}))\mathbf{o}(\mathbf{v})) \end{aligned}$$

Let an upper bound for $\|D\mathbf{h}^{-1}(\mathbf{h}(\mathbf{x}))\|$ be C . It exists because \bar{V} is compact and \mathbf{h}^{-1} is C^1 on an open set containing this compact set. Therefore, since all the boxes in \mathcal{B}_m are in diameter less than δ ,

$$\begin{aligned} \mathbf{h}(B(\mathbf{x}, r)) - \mathbf{h}(\mathbf{x}) &= \\ \mathbf{h}(\mathbf{x} + B(\mathbf{0}, r)) - \mathbf{h}(\mathbf{x}) &\subseteq D\mathbf{h}(\mathbf{x})(B(\mathbf{0}, (1 + C\varepsilon)r)). \end{aligned} \quad (33.8)$$

Then choose m still larger if necessary so that $f(\mathbf{y})$ is uniformly approximated by

$$\sum_i f(\mathbf{h}(\mathbf{x}_i^m)) \mathcal{X}_{\mathbf{h}(R_i^m)}(\mathbf{y}), \quad \mathbf{x}_i^m \in R_i^m,$$

to within ε . Let r_i^m be half the diameter of R_i^m . Thus $\sum_i m_p(B(\mathbf{0}, r_i^m)) = m_p(U)$. This is by the formula for the measure of a box. It is just the product of the lengths of the sides. Recall the norm is $\|\cdot\|_\infty$ so the balls are boxes.

$$m_p(R_i^m) = m_p(B(\mathbf{x}_i^m, r_i^m)) = m_p(B(\mathbf{0}, r_i^m))$$

Then

$$\begin{aligned} \int_V f(\mathbf{y}) dm_p &= \sum_{i=1}^\infty \int_{\mathbf{h}(R_i^m)} f(\mathbf{y}) dm_p \\ &\leq \varepsilon m_p(V) + \sum_{i=1}^\infty \int_{\mathbf{h}(R_i^m)} f(\mathbf{h}(\mathbf{x}_i^m)) dm_p \\ &\leq \varepsilon m_p(V) + \sum_{i=1}^\infty f(\mathbf{h}(\mathbf{x}_i^m)) m_p(\mathbf{h}(R_i^m)) \\ &\leq \varepsilon m_p(V) + \sum_{i=1}^\infty f(\mathbf{h}(\mathbf{x}_i^m)) m_p(D\mathbf{h}(\mathbf{x}_i^m)(B(\mathbf{0}, (1 + C\varepsilon)r_i))) \\ &= \varepsilon m_p(V) + (1 + C\varepsilon)^p \sum_{i=1}^\infty \int_{R_i^m} f(\mathbf{h}(\mathbf{x}_i^m)) |\det(D\mathbf{h}(\mathbf{x}_i^m))| dm_p \\ &\leq \varepsilon m_p(V) + (1 + C\varepsilon)^p \sum_{i=1}^\infty \left(\int_{R_i^m} f(\mathbf{h}(\mathbf{x})) |\det(D\mathbf{h}(\mathbf{x}))| dm_p + 2\varepsilon m_p(R_i^m) \right) \\ &\leq \varepsilon m_p(V) + (1 + C\varepsilon)^p \sum_{i=1}^\infty \int_{R_i^m} f(\mathbf{h}(\mathbf{x})) |\det(D\mathbf{h}(\mathbf{x}))| dm_p \\ &\quad + (1 + C\varepsilon)^p 2\varepsilon m_p(U) \end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, this shows

$$\int_V f(\mathbf{y}) dm_p \leq \int_U f(\mathbf{h}(\mathbf{x})) |\det(D\mathbf{h}(\mathbf{x}))| dm_p \quad (33.9)$$

whenever f is uniformly continuous and bounded on V . Now

$$\mathbf{x} \rightarrow f(\mathbf{h}(\mathbf{x})) |\det(D\mathbf{h}(\mathbf{x}))|$$

has the same properties as f and so, using the same argument with U and V switching roles and replacing \mathbf{h} with \mathbf{h}^{-1} ,

$$\begin{aligned} &\int_U f(\mathbf{h}(\mathbf{x})) |\det(D\mathbf{h}(\mathbf{x}))| dm_p \\ &\leq \int_V f(\mathbf{h}(\mathbf{h}^{-1}(\mathbf{y}))) |\det(D\mathbf{h}(\mathbf{h}^{-1}(\mathbf{y})))| |\det(D\mathbf{h}^{-1}(\mathbf{y}))| dm_p = \int_V f(\mathbf{y}) dm_p \end{aligned}$$

by the chain rule. This with 33.9 proves the lemma. ■

The Lebesgue integral is defined for nonnegative functions and then you break up an arbitrary function into positive and negative parts. Thus the most convenient theorems involve nonnegative functions which do not involve assumptions of uniform continuity and such things. The next corollary gives such a result. This will remove assumptions that U, V are bounded and the need for larger open sets on which $\mathbf{h}, \mathbf{h}^{-1}$ are defined and C^1 .

Corollary 33.6.2 *Let U be an open set in \mathbb{R}^p and let \mathbf{h} be a one to one C^1 function such that $\mathbf{h}(U) = V$ and $|\det D\mathbf{h}(\mathbf{x})| \neq 0$ for all \mathbf{x} . Let f be continuous and nonnegative defined on V . Then*

$$\int_V f(\mathbf{y}) dm_p = \int_U f(\mathbf{h}(\mathbf{x})) |\det(D\mathbf{h}(\mathbf{x}))| dm_p$$

Proof: Let $U_k \equiv (-k, k)^p \cap \{\mathbf{x} \in U : \text{dist}(\mathbf{x}, U^C) > \frac{1}{k}\}$. Thus $\overline{U_k} \subseteq U_{k+1}$ for all k and $\overline{U_k}$ is closed and bounded, hence compact. The inverse function theorem 24.0.5 implies $V_k \equiv \mathbf{h}(U_k)$ is open and \mathbf{h}^{-1} is C^1 on V_k . Also f is uniformly continuous on $\overline{U_k}$ hence on U_k as well. It follows that

$$\int_V \mathcal{X}_{\mathbf{h}(U_k)}(\mathbf{y}) f(\mathbf{y}) dm_p = \int_U \mathcal{X}_{U_k}(\mathbf{x}) f(\mathbf{h}(\mathbf{x})) |\det(D\mathbf{h}(\mathbf{x}))| dm_p$$

Now use the monotone convergence theorem and let $k \rightarrow \infty$. ■

It is easy to generalize this corollary to the case where f is nonnegative and only Borel measurable. Let $R = \prod_{i=1}^p (a_i, b_i)$, a open rectangle. Then let $g^k(\mathbf{x}) \equiv \prod_{i=1}^p g_i^k(x_i)$ where $g_i^k(t) \geq 0$, is continuous, piecewise linear, equals 0 off (a_i, b_i) and 1 on $[a_i + \frac{1}{k}, b_i - \frac{1}{k}]$. Thus $\lim_{k \rightarrow \infty} g^k(\mathbf{x}) = \mathcal{X}_R(\mathbf{x})$ and $g^k(\mathbf{x}) \leq g^{k+1}(\mathbf{x})$ for all k . Therefore, apply the monotone convergence theorem to obtain

$$\begin{aligned} \int_U \mathcal{X}_R(\mathbf{h}(\mathbf{x})) |\det(D\mathbf{h}(\mathbf{x}))| dm_p &= \lim_{k \rightarrow \infty} \int_U g^k(\mathbf{h}(\mathbf{x})) |\det(D\mathbf{h}(\mathbf{x}))| dm_p \\ &= \lim_{k \rightarrow \infty} \int_{\mathbf{h}(U)} g^k(\mathbf{y}) dm_p = \int_{\mathbf{h}(U)} \mathcal{X}_R(\mathbf{y}) dm_p \end{aligned}$$

Now let \mathcal{K} be the pi system of open rectangles. Thus $\sigma(\mathcal{K}) = \mathcal{B}(\mathbb{R}^p)$. Let $R_k \equiv \prod_{i=1}^p (-k, k)$

$$\mathcal{G} \equiv \left\{ E \in \mathcal{B}(\mathbb{R}^p) : \int_U \mathcal{X}_{E \cap R_k}(\mathbf{h}(\mathbf{x})) |\det(D\mathbf{h}(\mathbf{x}))| dm_p = \int_{\mathbf{h}(U)} \mathcal{X}_{E \cap R_k}(\mathbf{y}) dm_p \right\}$$

Then it is routine to verify that \mathcal{G} is closed with respect to countable disjoint unions and complements. The assertion about disjoint unions is obvious. Consider the one about complements. Say $E \in \mathcal{G}$. Then

$$\begin{aligned} \int_{\mathbf{h}(U)} \mathcal{X}_{E \cap R_k}(\mathbf{y}) dm_p + \int_{\mathbf{h}(U)} \mathcal{X}_{E^C \cap R_k}(\mathbf{y}) dm_p &= \int_{\mathbf{h}(U)} \mathcal{X}_{R_k}(\mathbf{y}) dm_p \\ &= \int_U \mathcal{X}_{R_k}(\mathbf{h}(\mathbf{x})) |\det(D\mathbf{h}(\mathbf{x}))| dm_p = \\ &\stackrel{A}{=} \int_U \mathcal{X}_{E \cap R_k}(\mathbf{h}(\mathbf{x})) |\det(D\mathbf{h}(\mathbf{x}))| dm_p + \int_U \mathcal{X}_{E^C \cap R_k}(\mathbf{h}(\mathbf{x})) |\det(D\mathbf{h}(\mathbf{x}))| dm_p \end{aligned}$$

It is known that $A = B$ so subtracting from both sides yields

$$\int_{\mathbf{h}(U)} \mathcal{X}_{E^c \cap R_k}(\mathbf{y}) dm_p = \int_U \mathcal{X}_{E^c \cap R_k}(\mathbf{h}(\mathbf{x})) |\det(D\mathbf{h}(\mathbf{x}))| dm_p$$

Therefore, by Dynkin's lemma, \mathcal{G} equals $\mathcal{B}(\mathbb{R}^p)$. Now let $k \rightarrow \infty$ and apply the monotone convergence theorem. The following theorem is now almost obvious because it was just shown that the change of variables formula holds for indicator functions of Borel sets and hence for every nonnegative simple function.

Theorem 33.6.3 *Let $f(\mathbf{y}) \geq 0$ and let it be Borel measurable. Also let \mathbf{h} be a one to one C^1 function on the open set U such that $\mathbf{h}(U) = V$ and $|\det D\mathbf{h}(\mathbf{x})| \neq 0$. Then*

$$\int_V f(\mathbf{y}) dm_p = \int_U f(\mathbf{h}(\mathbf{x})) |\det(D\mathbf{h}(\mathbf{x}))| dm_p$$

Proof: By Theorem 32.2.8, there exists an increasing sequence of Borel measurable simple functions $\{s_k\}$ which converges pointwise to $f(\mathbf{y})$. Then by the monotone convergence theorem,

$$\begin{aligned} \int_V f(\mathbf{y}) dm_p &= \lim_{k \rightarrow \infty} \int_V s_k(\mathbf{y}) dm_p = \lim_{k \rightarrow \infty} \int_U s_k(\mathbf{h}(\mathbf{x})) |\det(D\mathbf{h}(\mathbf{x}))| dm_p \\ &= \int_U f(\mathbf{h}(\mathbf{x})) |\det(D\mathbf{h}(\mathbf{x}))| dm_p \blacksquare \end{aligned}$$

A lot more can be done. See a more advanced book for these things. My on line book Calculus of real and complex variables has it. You don't need to have \mathbf{h} be C^1 . Differentiable is enough. You also don't need to assume $|\det D\mathbf{h}(\mathbf{x})| \neq 0$ or even that \mathbf{h} is differentiable everywhere, but this is a good beginning result.

33.7 Exercises

1. Show $\int_0^M \sin t \int_0^\infty e^{-xt} dx dt = \int_0^M \frac{\sin t}{t} dt$. Use Fubini's theorem to interchange the order of integration and eventually conclude that the improper Riemann integral $\int_0^\infty \frac{\sin t}{t} dt$ exists and is $\frac{1}{2}\pi$.
2. This problem will help to understand that a certain kind of function exists.

$$f(x) = \begin{cases} e^{-1/x^2} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

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.

3. \uparrow Let $f(x)$ be as given above. Now let

$$\hat{f}(x) \equiv \begin{cases} f(x) & \text{if } x \leq 0 \\ 0 & \text{if } x > 0 \end{cases}$$

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 $\psi(x) = \prod_{k=1}^p g(x_k)$. For $U = B(\mathbf{0}, 2r)$ with the norm given by

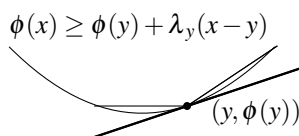
$$\|x\| = \max\{|x_k|, k \leq p\},$$

show that $\psi \in C_c^\infty(U)$.

4. \uparrow Using the above problem, let $\psi \in C_c^\infty(B(\mathbf{0}, 1))$. Also let $\psi \geq 0$ as in 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 $\psi_k(x) \equiv k^p \psi(kx)$. Show that ψ_k equals zero off a compact subset of $B(\mathbf{0}, \frac{1}{k})$ and $\int \psi_k dm_p = 1$. We say that $\text{spt}(\psi_k) \subseteq B(\mathbf{0}, \frac{1}{k})$. $\text{spt}(f)$ is defined as the closure of the set on which f is not equal to 0. Such a sequence of functions as just defined $\{\psi_k\}$ where $\int \psi_k dm_p = 1$ and $\psi_k \geq 0$ and $\text{spt}(\psi_k) \subseteq B(\mathbf{0}, \frac{1}{k})$ is called a **mollifier**.
5. If you have $f \in L^1(\mathbb{R}^p)$ with respect to Lebesgue measure and ψ_k is a mollifier, show that $f * \psi_k(x) \equiv \int_{\mathbb{R}^p} f(x-y) \psi_k(y) dm_p$ is infinitely differentiable. **Hint:** First show it equals $\int_{\mathbb{R}^p} f(y) \psi_k(x-y) dm_p$. Then use dominated convergence theorem.
6. Let $\phi : \mathbb{R} \rightarrow \mathbb{R}$ be convex. This means

$$\phi(\lambda x + (1-\lambda)y) \leq \lambda \phi(x) + (1-\lambda)\phi(y)$$

whenever $\lambda \in [0, 1]$. The following picture illustrates what is about to be shown.



- (a) Show that for $x < y < z$, $\frac{\phi(z)-\phi(x)}{z-x} \leq \frac{\phi(z)-\phi(y)}{z-y}$.
- (b) Next show $\frac{\phi(z)-\phi(x)}{z-x} \geq \frac{\phi(y)-\phi(x)}{y-x}$. To do these, use convexity applied to y .
- (c) Conclude $\frac{\phi(z)-\phi(y)}{z-y} \geq \frac{\phi(z)-\phi(x)}{z-x} \geq \frac{\phi(y)-\phi(x)}{y-x}$. In particular, $\frac{\phi(z)-\phi(y)}{z-y} \geq \frac{\phi(y)-\phi(x)}{y-x}$. (Difference quotients increase from left to right.)
- (d) Let $\lambda_y \equiv \inf \left\{ \frac{\phi(z)-\phi(y)}{z-y} : z > y \right\}$. Then for $y \geq x$,

$$\phi(y) - \phi(x) \leq \lambda_y(y-x)$$

Show that even if $x > y$, the same inequality holds. Thus for all x ,

$$\phi(y) + \lambda_y(x-y) \leq \phi(x)$$

- (e) Next show that for $x \in (y, z)$,

$$\phi(y) + \lambda_y(x-y) \leq \phi(x) \leq \phi(y) + a_z(x-y)$$

for some a_z and for $x \in (w, y)$, there is a_w such that

$$\phi(y) + \lambda_y(x-y) \leq \phi(x) \leq \phi(y) + a_w(x-y).$$

Thus ϕ is continuous. Note that there is no change if ϕ is convex on an open interval.

7. \uparrow Prove Jensen's inequality. If $\phi : \mathbb{R} \rightarrow \mathbb{R}$ is convex, $\mu(\Omega) = 1$, and $f : \Omega \rightarrow \mathbb{R}$ is in $L^1(\Omega)$, then $\phi(\int_{\Omega} f d\mu) \leq \int_{\Omega} \phi(f) d\mu$. **Hint:** Let $s = \int_{\Omega} f d\mu$ and use Problem 6.
8. $B(p, q) = \int_0^1 x^{p-1}(1-x)^{q-1} dx$, $\Gamma(p) = \int_0^{\infty} e^{-t} t^{p-1} dt$ 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.) $\Gamma(p)\Gamma(q) = B(p, q)\Gamma(p+q)$.
9. If $X : \Omega \rightarrow \mathbb{R}^p$ is measurable, where (Ω, \mathcal{F}) is a measurable space, show that

$$X^{-1}(B) \in \mathcal{F}$$

for every B a Borel set.

10. \uparrow Let (Ω, \mathcal{F}, P) be a probability space. This means the measure P has the property that $P(\Omega) = 1$. A random vector is a measurable function $X : \Omega \rightarrow \mathbb{R}^p$. The probability distribution measure λ_X is defined as follows. For E Borel, $\lambda_X(E) \equiv P(X \in E)$. Show this gives a probability measure on the Borel sets of \mathbb{R}^p . Sometimes this measure can be realized as an integral of the form $\int_{\mathbb{R}^p} f(x) dm_p$ but in general, this will not be the case. However, it is a perfectly good measure and all the theory of the Lebesgue integral developed above can be used.
11. \uparrow In the context of the above problem, suppose E is a Borel set in \mathbb{R}^p . Note first that the concept of a Borel set is not even defined on Ω . Explain the following equations:

$$\begin{aligned} \int \mathcal{X}_E(x) d\lambda_X &= \lambda_X(E) \equiv P(X \in E) = P(X^{-1}(E)) \\ &= \int \mathcal{X}_{X^{-1}(E)}(\omega) dP = \int \mathcal{X}_E(X(\omega)) dP \end{aligned}$$

Be sure to explain why everything makes sense and is appropriately measurable. Extend this to conclude that if f is a Borel measurable nonnegative function, then

$$\int f d\lambda_X = \int f(X(\omega)) dP$$

Hopefully you will see from this that there are lots of measures which are of interest other than Lebesgue measure.

12. Show that if f is any bounded Borel measurable function, then

$$\int f d\lambda_X = \int f(X(\omega)) dP$$

13. To integrate complex valued functions $f : \Omega \rightarrow \mathbb{C}$, first note that these are defined to be measurable if the real and imaginary parts are measurable. Then

$$\int f d\mu = \int (\operatorname{Re} f) d\mu + i \int (\operatorname{Im} f) d\mu$$

In the context of probability distribution measures described above, explain why everything makes sense and for $t \in \mathbb{R}^p$

$$\int_{\Omega} e^{iX(\omega) \cdot t} dP = \int_{\mathbb{R}^p} e^{ix \cdot t} d\lambda_X \equiv \phi_X(t)$$

This is called the characteristic function. It turns out that these completely characterize the probability distribution measures but this is a topic for a more advanced book which has important representation theorems not discussed here.

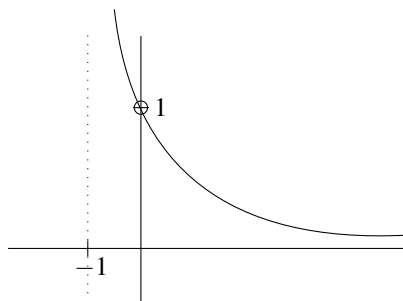
14. Show that there exists a subset of \mathbb{R} consisting of everything off a set of measure no more than ε which contains no intervals.
15. This problem outlines an approach to Stirling's formula following [26] and [8]. From the above problems, $\Gamma(n+1) = n!$ for $n \geq 0$. Consider more generally $\Gamma(x+1)$ for $x > 0$. Actually, we will always assume $x > 1$ since it is the limit as $x \rightarrow \infty$ which is of interest. $\Gamma(x+1) = \int_0^\infty e^{-t} t^x dt$. Change variables letting $t = x(1+u)$ to obtain

$$\Gamma(x+1) = x^{x+1} e^{-x} \int_{-1}^\infty ((1+u)e^{-u})^x du$$

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 $u \rightarrow 0$ 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 $u \rightarrow \infty$.



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 $s < 1$, $h\left(s\sqrt{\frac{2}{x}}\right) > 2 - 2\ln 2 = 0.61371$ so the above integrand is dominated by $e^{-(2-2\ln 2)s^2}$. Consider the integrand in the above for $s > 1$. The exponent part is

$$\begin{aligned} & -s^2 \left(\frac{2}{\left(s\sqrt{\frac{2}{x}}\right)^2} \left(s\sqrt{\frac{2}{x}} - \ln\left(1 + s\sqrt{\frac{2}{x}}\right) \right) \right) \\ &= -s^2 \left(\frac{\sqrt{2}}{s} \sqrt{x} - \frac{1}{s^2} x \ln\left(1 + s\sqrt{\frac{2}{x}}\right) \right) \end{aligned}$$

$$= - \left(\sqrt{2}\sqrt{x}s - x \ln \left(1 + s\sqrt{\frac{2}{x}} \right) \right)$$

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

$$\left(\sqrt{2}\sqrt{1}s - \ln \left(1 + s\sqrt{\frac{2}{1}} \right) \right)$$

and so

$$\begin{aligned} \exp \left(-s^2 h \left(s\sqrt{\frac{2}{x}} \right) \right) &\leq \exp \left(- \left(\sqrt{2}\sqrt{1}s - \ln \left(1 + s\sqrt{\frac{2}{1}} \right) \right) \right) \\ &= (1 + s\sqrt{2}) e^{-\sqrt{2}s} \end{aligned}$$

Thus, there exists a dominating function for $\mathcal{X}_{[-\sqrt{x/2}, \infty]}(s) \exp \left(-s^2 h \left(s\sqrt{\frac{2}{x}} \right) \right)$ and these functions converge pointwise to $\exp(-s^2)$ so by the dominated convergence theorem,

$$\lim_{x \rightarrow \infty} \int_{-\sqrt{x/2}}^{\infty} \exp \left(-s^2 h \left(s\sqrt{\frac{2}{x}} \right) \right) ds = \int_{-\infty}^{\infty} e^{-s^2} ds = \sqrt{\pi}$$

See Problem 10 on Page 249. This yields a general Stirling's formula,

$$\lim_{x \rightarrow \infty} \frac{\Gamma(x+1)}{x^x e^{-x} \sqrt{2x}} = \sqrt{\pi}.$$

16. This problem is on the Dirichlet integral which is $\int_0^{\infty} \frac{\sin x}{x} dx$. Show that the integrand is not in $L^1(0, \infty)$. However, verify that $\lim_{r \rightarrow \infty} \int_0^r \frac{\sin x}{x} dx$ exists and equals $\frac{\pi}{2}$. **Hint:** Explain why $\int_0^r \frac{\sin x}{x} dx = \int_0^r \sin x \int_0^{\infty} e^{-tx} dt dx$. Then use Fubini's theorem to write this last is equal to

$$\int_0^{\infty} \int_0^r \sin(x) e^{-tx} dx dt$$

Integrate by parts in the inside integral to obtain

$$\begin{aligned} &= \int_0^{\infty} \frac{1}{t^2 + 1} - e^{-rt} \left(\frac{\cos r}{1 + t^2} + t \frac{\sin r}{1 + t^2} \right) dt \\ &= \frac{\pi}{2} - \int_0^{\infty} \frac{1}{\sqrt{1 + t^2}} e^{-rt} \cos(r - \phi(t)) dt \end{aligned}$$

Explain why the second integral converges to 0 as $r \rightarrow \infty$.

Bibliography

- [1] **Apostol, T. M.**, *Calculus second edition*, Wiley, 1967.
- [2] **Apostol T.M.** *Calculus Volume II Second edition*, Wiley 1969.
- [3] **Apostol, T. M.**, *Mathematical Analysis*, Addison Wesley Publishing Co., 1974.
- [4] **Baker, R.**, *Linear Algebra*, Rinton Press 2001.
- [5] **Baker, R., Christenson, C., and Orde, H.**, *Collected papers / Bernhard Riemann ; translated from the 1892 edition by Roger Baker, Charles Christenson and Henry Orde*. Kendrick Press, 2004.
- [6] **Bartle R.G.**, *A Modern Theory of Integration*, Grad. Studies in Math., Amer. Math. Society, Providence, RI, 2000.
- [7] **Bartle R. G. and Sherbert D.R.** *Introduction to Real Analysis* third edition, Wiley 2000.
- [8] **Buck, R. C.** *Advanced Calculus* 2 edition. McGraw-Hill, 1965.
- [9] **Chahal J. S.** , *Historical Perspective of Mathematics* 2000 B.C. - 2000 A.D.
- [10] **Davis H. and Snider A.**, *Vector Analysis* Wm. C. Brown 1995.
- [11] **D'Angelo, J. and West D.** *Mathematical Thinking Problem Solving and Proofs*, Prentice Hall 1997.
- [12] **Edwards C.H.** *Advanced Calculus of several Variables*, Dover 1994.
- [13] **Euclid**, *The Thirteen Books of the Elements*, Dover, 1956.
- [14] **Eves, H.** *An Introduction To The History of Mathematics*, Holt Rinehart and Winston 1976.
- [15] **Fitzpatrick P. M.**, *Advanced Calculus a course in Mathematical Analysis*, PWS Publishing Company 1996.
- [16] **Fleming W.**, *Functions of Several Variables*, Springer Verlag 1976.
- [17] **Greenberg, M.** *Advanced Engineering Mathematics*, Second edition, Prentice Hall, 1998
- [18] **Gurtin M.** *An introduction to continuum mechanics*, Academic press 1981.

- [19] **Hardy G.**, *A Course Of Pure Mathematics, Tenth edition*, Cambridge University Press 1992.
- [20] **Hewitt E.** and **Stromberg K.** *Real and Abstract Analysis*, Springer-Verlag, New York, 1965.
- [21] **Kuttler K.**, *Calculus Theory and Applications*, Volumes 1 and 2., World Scientific. 2011.
- [22] **McLeod R.** *The Generalized Riemann Integral*, Mathematical Association of America, Carus Mathematical Monographs number 20 1980.
- [23] **McShane E. J.** *Integration*, Princeton University Press, Princeton, N.J. 1944.
- [24] **Nobel B. and Daniel J.** *Applied Linear Algebra*, Prentice Hall, 1977.
- [25] **Rose, David, A.**, The College Math Journal, vol. 22, No.2 March 1991.
- [26] **Rudin, W.**, *Principles of mathematical analysis*, McGraw Hill third edition 1976
- [27] **Rudin W.**, *Real and Complex Analysis*, third edition, McGraw-Hill, 1987.
- [28] **Salas S. and Hille E.**, *Calculus One and Several Variables*, Wiley 1990.
- [29] **Sears and Zemansky**, *University Physics, Third edition*, Addison Wesley 1963.
- [30] **Spivak M.**, *Calculus On Manifolds*, Benjamin 1965.
- [31] **Tierney, John**, *Calculus and Analytic Geometry*, fourth edition, Allyn and Bacon, Boston, 1969.
- [32] **Widder, D.** *Advanced Calculus*, second edition, Prentice Hall 1961.

Index

- C^1 , 486
- C^k , 486
- Δ , 588
- \cap , 21
- \cup , 21
- ∇^2 , 588
- π systems, 673
- n^{th} term test, 169

- absolute convergence, 166, 173
 - rearrangement, 166
- absolute value, 24
 - complex number, 41
- acceleration, 148, 203
- additive inverse
 - unique, 18
- adjugate, 433, 460
- agony, pain and suffering, 543
- alternating series, 172
- alternating series test, 172
- amplitude, 68
- angle between planes, 314
- angle between vectors, 297
- angles
 - degrees, 54
 - radian measure, 54
- angular velocity, 310
- angular velocity vector, 376
- annuity
 - ordinary, 32
- antiderivative, 181
- antiderivatives
 - integration by parts, 205
 - partial fractions, 214
 - tabular integration, 206
- arc length, 350
- Archimedian property, 30
- area
 - circular sector, 65
 - area between two graphs, 209
 - area of a parallelogram, 304
- areas
 - surfaces of revolution, 228
- arithmetic mean, 520
- augmented matrix, 401
- auspicious substitutions, 213
- average velocity, 148

- balance of momentum, 600
- basis, 417
- basis, 633
- Bernoulli law, 496
- Bernstein polynomials, 118
- Bezier curves, 357
- binomial series, 258
- binomial theorem, 28, 33
 - infinite series, 260
- binormal, 363
- Borel sets, 674
- bounded, 97, 325
- box product, 306

- Caratheodory's procedure, 662
- cardioid, 273
- Cartesian coordinates, 282
- Cartesian product, 52, 324
- Cauchy, 104
- Cauchy condensation test, 168
- Cauchy criterion for sums, 165
- Cauchy mean value theorem, 147
- Cauchy product, 175
- Cauchy Schwarz inequality, 290, 296
- Cauchy sequence, 98
 - convergence, 99
- Cauchy stress, 603
- Cavendish, 371
- Cayley Hamilton theorem, 460
- center of mass, 309

- central force, 365
- central force field, 370
- centripetal force, 276
- chain rule, 132
 - functions of many variables, 490
- change of variables formula, 186, 571
- characteristic polynomial, 435, 460
- chi-squared, 251
- circular functions, 76
- circulation density, 630
- classical adjoint, 433
- closed and bounded
 - sequentially compact, 97
- closed set, 322
- closed subset of compact set, 97
- coefficient of thermal conductivity, 502
- cofactor, 426, 428, 456
- cofactor matrix, 428
- column space, 415
- compact, 341
- compactness
 - preservation, 110
 - continuous function, 110
- comparison test, 166, 167
- complement, 322
- complement of a set, 95
- completeness, 34
 - convergence of Cauchy sequence, 99
- completeness axiom, 35
- completing the square, 37
- complex conjugate, 41
- complex numbers, 40
 - roots, 44
- complex numbers
 - arithmetic, 40
 - triangle inequality, 42
- component, 302
- component of a force, 300
- conditional convergence, 166
- conjugate
 - of a product, 47
- conservation of mass, 600
- conservative, 625
 - path independent line integral, 536
 - vector field, 536
- constitutive laws, 606
- continuity
 - equivalent formulations, 106
- inverse function, 112
- limits of sequences, 106
- litany of properties, 106
- on a compact set, 114
- one to one, 111
- preservation of inequality, 106
- uniform, 114
- continuous
 - at one point, 105
- continuous and one to one
 - monotone, 111
- continuous function, 104
 - only at irrationals, 108
- continuous functions
 - combinations, 106
 - supremum and infimum, 121
- continuous image of compact set, 110
- contitional convergence, 173
- contour graph, 467
- convergence
 - pointwise, 336
 - uniform, 336
- Coordinates, 281
- Coriolis acceleration
 - earth, 381
- Coriolis force, 276
- countable, 652
- Cramer's rule, 462
- critical point, 506
- critical points, 141
- cross product, 303
 - area of parallelogram, 304
 - coordinate description, 304
 - geometric description, 303
- curl, 587
- curvature, 359, 363
 - independence, 360
- cycloid, 630
- D'Alembert, 476
- Darboux integral, 190
- De Moivre theorem, 46
- De Moivre's theorem, 43
- deformation gradient, 601
- degree, 55
- DeMorgan's laws, 21
- dense, 31
- density and mass, 544

- density of rationals, 30
- dependent, 416
- derivative, 479, 480
 - chain rule, 132
 - equals zero, then function is constant, 148
 - equivalent difference quotient, 132
 - higher order derivatives, 133
 - intermediate value property, 149
 - inverse function, 133
 - mean value theorem, 147
 - product rule, 132
 - quotient rule, 133
 - sum, product, quotient, chain rule, 132
- derivative of a function, 344
- derived series, 255
- determinant, 451
 - alternating property, 454
 - cofactor, 426
 - cofactor expansion, 456
 - expanding along row or column, 427
 - expansion along row (column), 456
 - matrix inverse formula, 433, 459
 - minor, 426
 - product, 430, 455
 - row operations, 429
 - transpose, 453
- determinant rank
 - row rank, 458
- difference quotient, 344
- differentiable, 477, 480
- differential equations, 641
- differentiation rules, 132, 347
- dimension, 417
- directed line segment, 286
- direction vector, 286
- directional derivative, 469, 500
- Dirichlet function, 50
- Dirichlet integral, 695
- Dirichlet test, 172
- discriminant, 46
- distance, 56, 207
- distance formula, 288
- divergence, 587
- divergence, 646
 - general curvilinear coordinates, 647
- divergence theorem, 592
- domain, 49
- dominated convergence theorem, 671
- donut, 580
- dot product, 295
 - geometric description, 297
- double series
 - absolute convergence, 175
 - interchange order of summation, 175
- double sum
 - interchange of order, 175
- dual basis, 634
- dual basis, 638
- Dynkin's lemma, 673
- eigenvalue, 519
- eigenvalues, 460
- eigenvectors, 424
- Einstein summation convention, 312
- elementary matrices, 397
- elementary matrix, 419
 - inverse, 400
 - properties, 400
- epigraph, 121
- equal area rule, 372
- equality of mixed partial derivatives, 474
- Euclidean algorithm, 31
- Euler's number, 75
- Eulerian coordinates, 601
- exponential growth, 242
- exponential growth and decay, 77
- extreme value theorem, 109
- Fatou's lemma, 669
- Fibonacci sequence, 84
- Fick's law, 502, 612
- field axioms, 18, 40
- finite intersection property, 99
 - compact sets, 100
- first derivative test, 150
- focus, 293
- force
 - on a dam, 232
- force field, 370, 533
- Foucault pendulum, 381
- Fourier law of heat conduction, 502
- Frenet Serret formulas, 364
- Frobinius norm, 446
- frustum of a cone, 228
- Fubini's theorem, 675

- Fubini's theorem, 248
- function
 - even, 144
 - odd, 144
 - uniformly continuous, 114
- fundamental matrix, 632
 - group property, 632
- fundamental theorem line integrals, 625
- fundamental theorem of algebra, 45, 277, 337
- fundamental theorem of algebra
 - plausibility argument, 277
- future value of an annuity, 32
- Gamma function, 693
- gamma function
 - existence and convergence, 240
 - properties, 241
- Gauss's theorem, 592
- geometric mean, 520
- geometric series, 164
 - sum, 32
- geometric series, 164
- gradient, 470
- gradient
 - contravariant components, 646
 - covariant components, 646
- Gram Schmidt process, 438
- graph of function, 52
- greatest lower bound, 35
- Green's theorem, 615, 625
- gronwall's inequality, 196
- half life, 79
- hanging chain, 385
- harmonic, 474
- heat equation, 474
- Heine Borel, 94
- Heine Borel theorem, 341
- Hessian matrix, 508, 522
- Holder continuous, 339
- Holder's inequality, 198
- homogeneous coordinates, 414
- hyperbolic functions, 76
- image, 415
- implicit differentiation, 139, 140
- implicit function theorem, 527, 530
- improper integral, 247
- improper Riemann integral, 240
- inconsistent, 396
- increment of volume
 - increment of area, 571
- independent, 416
- index
 - lowering, 635
 - raising, 635
- infinite series
 - raised to a power, 266
- infinite sums
 - properties, 164
- inner product, 295
- integral
 - continuous function, 185
 - decreasing function, 665
 - definition, 190
 - uniform convergence, 188
- integration by parts, 186, 205
- intercepts, 316
- interest
 - compounded continuously, 159
- interior point, 95, 322
- intermediate value theorem, 110, 111
- interval of convergence, 254
- inverse, 406
 - left inverse, 460
 - right inverse, 460
- inverse function theorem, 530, 531
- inverse image, 51
- inverses and determinants, 459
- invertible, 406
- iterated integral, 539, 685
- iterated integrals, 248, 684
- Jacobian determinant, 571
- Jensens inequality, 693
- joule, 301
- Kepler's first law, 372
- Kepler's laws, 371
- Kepler's third law, 374
- kilogram, 309
- kinetic energy, 640
- Kroneker delta, 312
- L'Hopital's rule, 155
- L'Hopitals rule, 157
- Lagrange multipliers, 517, 531, 532

- Lagrange remainder, 152, 153, 522
- Lagrangian coordinates, 601
- Lagrangian formalism, 641
- Laplace expansion, 456
- Laplace transform, 243
 - obvious properties, 243
- Laplacian, 474
 - polar coordinates, 500
- Laplacian
 - general curvilinear coordinates, 647
- law of cosines, 61
- least squares regression, 475
- least upper bound, 35
- Lebesgue integral
 - desires to be linear, 670
 - nonnegative function, 666
 - simple function, 668
- Lebesgue number, 341
- length of smooth curve, 351
- Leontief model, 413
- \liminf , 120
- \limsup , 120
- \liminf , 90
- limit comparison test, 167
- limit of a function, 329
- limit of a subsequence, 90
- limit of n th root of n , 253
- limit point, 96, 465
- limit points
 - closed sets, 97
- limitfore, 93
- limits
 - at infinity, 121
 - properties, 123
 - squeezing theorem, 88, 124
 - uniqueness, 86, 122
 - well defined, 122
- limits and continuity, 331
- limits of sequences
 - preservation of order, 90
 - properties, 87
- \limsup , 90
- line integral, 534
- linear combination, 390, 414, 454
- linear functions, 389
- linear initial value problem, 195
- linear map, 389
- linear transformation, 389
- lines
 - parametric equation, 286
- Lipschitz, 120, 338, 339
- lizards
 - surface area, 578
- local extrema, 141
- local extremum, 505
 - derivative equals 0, 141
- local maximum, 141, 505
- local minimum, 141, 505
- locating local extremum, 506
- logistic function, 79
- Lotka Volterra equations, 334
- lower semicontinuous, 120
- lower sums, 71, 188
- main diagonal, 429
- mass balance, 600
- material coordinates, 601
- mathematical induction, 29
- matrices
 - eigenvalues exist, 435
- matrix
 - left inverse, 460
 - lower triangular, 428
 - right inverse, 460
 - upper triangular, 428
- maximizing sequence, 120
- mean value theorem
 - Cauchy, 147
 - usual version, 147
- measurability
 - limit of simple functions, 658
- measurable, 661
- measurable sets, 661
- measurable space, 655
- measure, 655
 - properties, 656
- measure space, 655
- measures
 - decreasing sequences of sets, 656
 - increasing sequences of sets, 656
- measures from outer measures, 662
- Merten's theorem, 176, 266
- metric tensor, 579, 634, 636
- metric tensor, 645
- minimal polynomial, 424
- minimizing sequence, 120

- minor, 426, 428, 456
- mixed partial derivatives, 473
- moment of a force, 308
- monic
 - polynomial, 39
- monotone convergence theorem, 669
- motion, 601
- multi-index, 329
- multiplicative inverse
 - unique, 18
- Navier, 613
- nested interval lemma, 92
- Neuman series, 413
- Newton Raphson procedure, 155
- Newton's second law, 641
- nonremovable discontinuity, 104
- normal vector to plane, 314
- one to one
 - rank, 423
- open set, 95, 322
- order, 23
- ordered fields, 23
- orientable, 622, 623
- orientation, 533
- oriented curve, 533
- origin, 281
- orthogonal matrix, 437
- orthonormal, 438
- osculating plane, 359, 362
- outer measure
 - measurable, 661
 - on \mathbb{R} , 681
- p series, 169
- parallelepiped
 - definition, 306
 - volume, 306, 462
- parameter, 286
- parametric curve, 343
- parametric equation, 286
- parametric function, 277
- parametrization, 76, 350
 - space curve, 343
- partial derivative, 470
- partial derivatives, 249
- partial fractions
 - expansion, 48
 - rules, 218
 - theory, 48
- partial summation formula, 172
- Pascal
 - triangle, 33
- permutation, 451
- permutation matrices, 397
- permutation symbol, 312
- permutations, 27
- perpendicular, 298
- phase shift, 68
- pi systems, 673
- piecewise continuous, 186
- Piola Kirchhoff stress, 605
- pivot columns, 406
- planes, 314
- pointwise convergence, 115
 - series, 178
- polar coordinates, 497
- polar form complex number, 43
- polynomial, 38, 51
 - addition, 38
 - degree, 38
 - division, 38
 - equality, 38
 - monic, 39
 - multiplication, 38
- polynomial
 - leading term, 38
 - monic, 38
- polynomials
 - factoring, 44
 - greatest common divisor, 39
- polynomials in n variables, 329
- position vector, 284
- postive number raised to real power, 76
- power series, 253
 - multiplication, 262
 - of a quotient, 268
 - raising to a power, 267
- present value of an annuity, 33
- preservation of compactness, 110
- prime numbers less given number, 178
- principal normal, 359, 363
- product measure, 675
- product of Borel sets, 674
- product rule, 132
 - cross product, 347

- dot product, 347
- projection, 420
- projection of a vector, 300
- Pythagorean theorem, 55
- quotient rule, 133
- radius of convergence, 254
- radius of curvature, 359, 362
- range, 49
- rank, 406, 419
- rank of a matrix, 457
- ratio test, 169
- rational function, 51
- rational function of cosines and sines, 220
- rational functions
 - sines and cosines, 220
- rational numbers, 17
 - dense, 31
 - density, 30
- real numbers, 17
- rearranged series
 - convergence, 166
- recurrence relation, 83
- recursively defined sequence, 83
- regular Sturm Liouville problem, 236
- related rates, 139
- relations
 - graph, 53
- removable discontinuity, 103
- Rieman integrable
 - functions of, 193
- Riemann integrable, 541, 546
- Riemann integral, 541, 546
- Riemann integral and Lebesgue integral, 683
- Riemann sum, 195
- Riemann sums, 541
- right handed system, 303
- Rolle's theorem, 147
- root
 - polynomial, 38
- root test, 170
- roots
 - existence, 36
- rot, 587
- rotation matrix, 424
- row operations, 397, 430
- row reduced echelon form, 403
- existence, 404
- unique, 405
- saddle point, 508
- scalar field, 587
- scalar multiplication, 283
- scalar potential, 625
- scalar product, 295
- scalars, 283
- second derivative test, 150, 523
- sequence of partial sums, 163
- sequences, 83
- sequentially compact
 - closed and bounded, 97
- series
 - absolute convergence, 166
 - conditional convergence, 166
 - convergence criterion, 165
 - double sum, 174
 - meaning of convergence, 163
 - multiplication of series, 175
 - nonnegative terms, 164
 - p series test, 168
 - ratio test, 169
 - root test, 170
- series of functions
 - uniform convergence, 178
- set
 - complement, 95
- sgn, 449
 - uniqueness, 451
- sigma algebra, 655
- sigma finite, 675
- sign of a permutation, 451
- singular point, 506
- skew symmetric, 395, 410
- smooth and not analytic, 162
- smooth curve, 350
- smoothness more general than analytic, 161
- spacial coordinates, 601
- span, 414, 454
- speed, 148
- spherical coordinates, 639
- squeezing theorem, 88
- step function, 187
- Stirling's formula, 240, 695
- Stoke's theorem, 620
- Stokes, 613

- subspace, 415
- subtend, 54
- summation notation, 19
- sup
 - increasing sequence, 163
 - interchange of order, 174
- symmetric, 395, 410
- symmetric form of a line, 287, 288
- symmetric matrices
 - diagonalization, 439
- tangent plane, 502
- Taylor polynomial
 - sine, 153
- Taylor series, 254
 - coefficients, 257
 - convergence and divergence, 254
 - differentiation, 255
 - multiplication, 262
 - of quotient, 268
 - raising to a power, 267
 - uniqueness, 257
- Taylor's formula, 152, 224, 522
- torque vector, 308
- torsion, 363
- torus, 580
- trace, 446
- traces, 316
- transformation rules, 642
- transpose of a matrix, 395
- triangle inequality, 24, 291, 296
 - complex numbers, 42
- trichotomy, 23
- trigonometric functions, 57
- trigonometric substitutions, 209
- uniform continuity, 335
- uniform convergence, 116
 - series, 178
- uniform norm, 336
- uniformly Cauchy
 - sequence of functions, 117
- uniformly continuous, 114, 247, 339, 341
- unit tangent vector, 359, 363
- upper semicontinuous, 120
- upper sums, 71, 188
- Urysohn's lemma, 342
- vector
 - scalar multiplication, 283
 - vector addition, 283
- vector
 - contravariant components, 635
 - covariant components, 635
- vector field, 533, 587
- vector fields, 332
- vector identities, 312
- vector potential, 589
- vector space axioms, 393
- vector valued function
 - continuity, 328
 - derivative, 344
 - integral, 344
 - limit theorems, 330
- vector valued functions, 327
- vectors, 281
- velocity, 148, 203
- volume
 - parallelepiped, 462
- volume element, 571
- volume increment, 571
- volume of parallelepiped, 569
- volume of unit ball in n dimensions, 610
- volumes
 - cross section, 183
- Wallis formula, 239
- wave equation, 474
- Weierstrass approximation
 - estimate, 117
- Weierstrass approximation theorem, 119
- Weierstrass Bolzano theorem, 337
- Weierstrass M test, 178
- well ordered, 28, 29
- work, 232, 534
- zero
 - polynomial, 38